

Accountability and Educational Improvement

Arnoud Oude Groote Beverborg
Tobias Feldhoff
Katharina Maag Merki
Falk Radisch *Editors*

Concept and Design Developments in School Improvement Research

Longitudinal, Multilevel and Mixed
Methods and Their Relevance for
Educational Accountability

OPEN ACCESS

 Springer

Accountability and Educational Improvement

Series Editors

Melanie Ehren, UCL Institute of Education, University College London,
London, UK

Katharina Maag Merki, Institut für Erziehungswissenschaft, Universität Zürich,
Zürich, Switzerland

This book series intends to bring together an array of theoretical and empirical research into accountability systems, external and internal evaluation, educational improvement, and their impact on teaching, learning and achievement of the students in a multilevel context. The series will address how different types of accountability and evaluation systems (e.g. school inspections, test-based accountability, merit pay, internal evaluations, peer review) have an impact (both intended and unintended) on educational improvement, particularly of education systems, schools, and teachers. The series addresses questions on the impact of different types of evaluation and accountability systems on equal opportunities in education, school improvement and teaching and learning in the classroom, and methods to study these questions. Theoretical foundations of educational improvement, accountability and evaluation systems will specifically be addressed (e.g. principal-agent theory, rational choice theory, cybernetics, goal setting theory, institutionalisation) to enhance our understanding of the mechanisms and processes underlying improvement through different types of (both external and internal) evaluation and accountability systems, and the context in which different types of evaluation are effective. These topics will be relevant for researchers studying the effects of such systems as well as for both practitioners and policy-makers who are in charge of the design of evaluation systems.

More information about this series at <http://www.springer.com/series/13537>

Arnoud Oude Groote Beverborg
Tobias Feldhoff • Katharina Maag Merki
Falk Radisch
Editors

Concept and Design Developments in School Improvement Research

Longitudinal, Multilevel and Mixed Methods
and Their Relevance for Educational
Accountability

 Springer

 ZENTRUM FÜR
SCHUL-, BILDUNGS-
UND HOCHSCHULFORSCHUNG

Editors

Arnoud Oude Groote Beverborg
Public Administration
Radboud University Nijmegen
Nijmegen, The Netherlands

Tobias Feldhoff
Institut für Erziehungswissenschaft
Johannes Gutenberg Universität
Mainz, Rheinland-Pfalz, Germany

Katharina Maag Merki
University of Zurich
Zurich, Switzerland

Falk Radisch
Institut für Schulpädagogik und Bildung
Universität Rostock
Rostock, Germany

This publication was supported by Center for School, Education, and Higher Education Research.



ISSN 2509-3320 ISSN 2509-3339 (electronic)
Accountability and Educational Improvement
ISBN 978-3-030-69344-2 ISBN 978-3-030-69345-9 (eBook)
<https://doi.org/10.1007/978-3-030-69345-9>

© The Editor(s) (if applicable) and The Author(s) 2021. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

1	Introduction	1
	Arnoud Oude Groote Beverborg, Tobias Feldhoff, Katharina Maag Merki, and Falk Radisch	
2	Why Must Everything Be So Complicated? Demands and Challenges on Methods for Analyzing School Improvement Processes	9
	Tobias Feldhoff and Falk Radisch	
3	School Improvement Capacity – A Review and a Reconceptualization from the Perspectives of Educational Effectiveness and Educational Policy	27
	David Reynolds and Annemarie Neeleman	
4	The Relationship Between Teacher Professional Community and Participative Decision-Making in Schools in 22 European Countries	41
	Catalina Lomos	
5	New Ways of Dealing with Lacking Measurement Invariance	63
	Markus Sauerwein and Désirée Theis	
6	Taking Composition and Similarity Effects into Account: Theoretical and Methodological Suggestions for Analyses of Nested School Data in School Improvement Research	83
	Kai Schudel and Katharina Maag Merki	
7	Reframing Educational Leadership Research in the Twenty-First Century	107
	David NG	

8	The Structure of Leadership Language: Rhetorical and Linguistic Methods for Studying School Improvement	137
	Rebecca Lowenhaupt	
9	Designing and Piloting a Leadership Daily Practice Log: Using Logs to Study the Practice of Leadership	155
	James P. Spillane and Anita Zuberi	
10	Learning in Collaboration: Exploring Processes and Outcomes	197
	B��n��dicte Vanblaere and Geert Devos	
11	Recurrence Quantification Analysis as a Methodological Innovation for School Improvement Research	219
	Arnoud Oude Groote Beverborg, Maarten Wijnants, Peter J. C. Slegers, and Tobias Feldhoff	
12	Regulation Activities of Teachers in Secondary Schools: Development of a Theoretical Framework and Exploratory Analyses in Four Secondary Schools Based on Time Sampling Data	257
	Katharina Maag Merki, Urs Grob, Beat Rechsteiner, Andrea Wullschleger, Nathanael Schori, and Ariane Rickenbacher	
13	Concept and Design Developments in School Improvement Research: General Discussion and Outlook for Further Research	303
	Tobias Feldhoff, Katharina Maag Merki, Arnoud Oude Groote Beverborg, and Falk Radisch	

About the Editors

Arnoud Oude Groote Beverborg currently works at the Department of Public Administration of the Nijmegen School of Management at Radboud University Nijmegen, the Netherlands. He worked as a post-doc at the Department of Educational Research and the Centre for School Effectiveness and School Improvement of the Johannes Gutenberg-University of Mainz, Germany, to which he is still affiliated. His work concentrates on theoretical and methodological developments regarding the longitudinal and reciprocal relations between professional learning activities, psychological states, workplace conditions, leadership, and governance. Additional to his interest in enhancing school change capacity, he is developing dynamic conceptualizations and operationalizations of workplace and organizational learning, for which he explores the application of dynamic systems modeling techniques.

Tobias Feldhoff is full professor for education science. He is head of the Center for School Improvement and School Effectiveness Research and chair of the Center for Research on School, Education and Higher Education (ZSBH) at the Johannes Gutenberg University Mainz. He is also co-coordinator of the Special Interest Group Educational Effectiveness and Improvement of the European Association for Research on Learning and Instruction (EARLI). His research topics are school improvement, school effectiveness, educational governance, and the link between them. One focus of his work is to develop designs and find methods to better understand school improvement processes, their dynamics, and effects. He is also interested in an organisation-theoretical foundation of school improvement.

Katharina Maag Merki is a full professor of educational science at the University of Zurich, Switzerland. Maag Merki's main research interests include research on school improvement, educational effectiveness, and self-regulated learning. She has over 20 years of experience in conducting complex interdisciplinary longitudinal analyses. Her research has been distinguished by several national and international grants. Her paper on "Conducting intervention studies on school improvement," published in the *Journal of Educational Administration*, was selected by the journal's editorial team as a Highly Commended Paper of 2014. At the moment, she is conducting a four-year multimethod longitudinal study to investigate mechanisms and effects of school improvement capacity on student learning in 60 primary schools in Switzerland. She is member of the National Research Council of the Swiss National Science Foundation.

Falk Radisch is an expert in research methods for educational research, especially school effectiveness, school improvement, and all day schooling. He has huge experience in planning, implementing, and analyzing large-scale and longitudinal studies. For his research, he has been using data sets from large-scale assessments like PISA, PIRLS, and TIMSS, as well as implementing large-scale and longitudinal studies in different areas of school-based research. He has been working on methodological problems of school-based research, especially for longitudinal, hierarchical, and non-linear methods for school effectiveness and school improvement research.

Chapter 1

Introduction



**Arnoud Oude Groote Beverborg, Tobias Feldhoff, Katharina Maag Merki,
and Falk Radisch**

Schools are continuously confronted with various forms of change, including changes in students' demographics, large-scale educational reforms, and accountability policies aimed at improving the quality of education. On the part of the schools, this requires sustained adaptation to, and co-development with, such changes to maintain or improve educational quality. As schools are multilevel, complex, and dynamic organizations, many conditions, factors, actors, and practices, as well as the (loosely coupled) interplay between them, can be involved therein (e.g. professional learning communities, accountability systems, leadership, instruction, stakeholders, etc.). School improvement can thus be understood through theories that are based on knowledge of systematic mechanisms that lead to effective schooling in combination with knowledge of context and path dependencies in individual school improvement journeys. Moreover, because theory-building, measuring, and analysing co-develop, fully understanding the school improvement process requires basic knowledge of the latest methodological and analytical developments and corresponding conceptualizations, as well as a continuous discourse on the link between theory and methodology. The complexity places high demands on the designs and methodologies from those who are tasked with empirically assessing and fostering improvements (e.g. educational researchers, quality care departments, and educational inspectorates).

A. Oude Groote Beverborg (✉)
Radboud University Nijmegen, Nijmegen, The Netherlands
e-mail: a.oudegrootebeverborg@fm.ru.nl

T. Feldhoff
Johannes Gutenberg University, Mainz, Germany

K. Maag Merki
University of Zurich, Zurich, Switzerland

F. Radisch
University of Rostock, Rostock, Germany

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research, Accountability and Educational Improvement*, https://doi.org/10.1007/978-3-030-69345-9_1

Traditionally, school improvement processes have been assessed with case studies. Case studies have the benefit that they only have to handle complexity within one case at a time. Complexity can then be assessed in a situated, flexible, and relatively easy way. Findings from case studies can also readily inform practice in those schools the studies were conducted in. However, case studies typically describe one specific example and do not test the mechanisms of the process, and therefore their findings cannot be generalized. As generalizability is highly valued, demands for designs and methodologies that can yield generalizable findings have been increasing within the fields of school improvement and accountability research. In contrast to case studies, quantitative studies are typically geared towards testing mechanisms and generalization. As such, quantitative studies are increasingly being conducted. Nevertheless, measurement and analysis of all aspects involved in improvement processes within and over schools and over time would be unfeasible in terms of the amount of measurement measures, the magnitude of the sample size, and the burden on the part of the participants. Thus, by assessing school improvement processes quantitatively, some complexity is necessarily lost, and therefore the findings of quantitative studies are also restricted.

Concurrent with the development towards a broader range of designs, the knowledge base has also expanded, and more sophisticated questions concerning the mechanisms of school improvement are being asked. This differentiation has led to a need for a discourse on how which available designs and methodologies can be aligned with which research questions that are asked in school improvement and accountability research. In our point of view the potential of combining the depth of case studies with the breadth of quantitative measurements and analyses in mixed-methods designs seems very promising; equally promising seems the adaptation of methodologies from related disciplines (e.g. sociology, psychology). Furthermore, application of sophisticated methodologies and designs that are sensitive to differences between contexts and change over time are needed to adequately address school improvement as a situated process.

With the book, we seek to host discussion of challenges in school improvement research and of methodologies that have the potential to foster school improvement research. Consequently, the focus of the book lies on innovative methodologies. As theory and methodology have a reciprocal relationship, innovative conceptualizations of school improvement that can foster innovative school improvement research will also be part of the book. The methodological and conceptual developments are presented as specific research examples on different areas of school improvement. In this way, the ideas, the chances, and the challenges can be understood in the context of the whole of each study, which, we think, will make it easier to apply these innovations and to avoid their pitfalls.

1.1 Overview of the Chapters

The chapters in this book give examples of the use of Measurement Invariance (in Structural Equation Models) to assess contextual differences (Chaps. 4 and 5), the Group Actor Partnership Interdependence Model and Social Network Analysis to

assess group composition effects (Chaps. 6 and 7, respectively), Rhetorical Analysis to assess persuasion (Chap. 8), logs as a measurement instrument that is sensitive to differences between contexts and change over time (Chaps. 9, 10, 11 and 12), Mixed Methods to show how different measurements and analyses can complement each other (Chap. 10), and Categorical Recurrence Quantification Analysis of the analysis of temporal (rather than spatial or causal) structures (Chap. 11). These innovative methodologies are applied to assess the following themes: complexity (Chaps. 2 and 7), context (Chaps. 3, 4, 5 and 6), leadership (Chaps. 7, 8 and 9), and learning and learning communities (Chaps. 4 and 10, 11 and 12).

In Chap. 2, Feldhoff and Radisch present a conceptualization of complexity in school improvement research. This conceptualization aims to foster understanding and identification of strengths, and possible weaknesses, of methodologies and designs. The conceptualization applies to both existing methodologies and designs as well as developments therein, such as those described in the studies in this book. More specifically, the chapter can be used by those who are tasked with empirically assessing and fostering improvements (e.g. educational researchers, departments of education, and educational inspectorates) to chart the demands and challenges that come with certain methodologies and designs, and to consider the focus and omissions of certain methodologies and designs when trying to answer research questions pertaining to specific aspects of the complexity of school improvement. This chapter is used in the last chapter to order the discussion of the other chapters.

In Chap. 3, Reynolds and Neeleman elaborate on the complexity of school improvement by discussing contextual aspects that need to be more extensively considered in research. They argue that there is a gap between research objects from educational effectiveness research on the one hand, and their incorporation into educational practice on the other hand. Central to their explanation of this gap is the neglect to account for the many contextual differences that can exist between and within schools (ranging from school leaders' values to student population characteristics), which resulted from a focus on 'what universally works'. The authors suggest that school improvement (research) would benefit from developments towards more differentiation between contexts.

In Chap. 4, Lomos presents a thorough example of how differences between contexts can be assessed. The study is concerned with differences between countries in how teacher professional community and participative decision-making are correlated. The cross-sectional questionnaire data from more than 35,000 teachers in 22 European countries in this study come from the International Civic and Citizenship Education Study (ICCS) 2009. The originality of the study lies in the assessment of how comparable the constructs are and how this affects the correlations between them. This is done by comparing the correlations between constructs based upon Exploratory Factor Analysis (EFA) with those based upon Multiple-Group Confirmatory Factor Analysis (MGCFA). In comparison to EFA, MGCFA includes the testing of measurement invariance of the latent variables between countries. Measurement invariance is seldom made the subject of discussion, but it is an important prerequisite in group (or time-point) comparisons, as it corrects for bias due to differences in understanding of constructs in different groups (or at different

time-points), and its absence may indicate that constructs have different meanings in different contexts (or that their meaning changes over time). The findings of the study show measurement invariance between all countries and higher correlations when constructs were corrected to have that measurement invariance.

In Chap. 5, Sauerwein and Theis use measurement invariance in the assessment of differences in the effects of disciplinary climate on reading scores between countries. This study is original in two ways. First, the authors show the false conclusions that the absence of measurement invariance may lead to, but second, they also show how measurement invariance, as a result in and of itself, may be explained by another variable that has measurement invariance (here: class size). The cross-sectional data from more than 20,000 students in 4 countries in this study come from the Programme for International Student Assessment (PISA) study 2009. Analysis of Variance (ANOVA) was used to assess the magnitude of the differences between countries in disciplinary climate and Regression Analysis was used to assess the effect of disciplinary climate on reading scores and of class size on disciplinary climate. As in Chap. 4, this was done twice: first without assessment of measurement invariance and then including assessment of measurement invariance. The findings of the study show that some comparisons of the magnitude of the differences in disciplinary climate and effect size between countries were invalid, due to the absence of measurement invariance there. Moreover, the authors assessed how patterns in how class size affected disciplinary climate resembled the patterns of the differences in measurement invariance in disciplinary climate between countries. They found that the effect of class size on disciplinary climate varied in accord with the differences in measurement invariance between countries. This procedure could uncover explanations of why the meaning of constructs differs between contexts (or time-points).

In contrast to the previous two chapters that focussed on between-group comparisons, in Chap. 6, Schudel and Maag Merki focus on within-group composition. They use the concept of diversity and assess the effect of staff members' positions within their teams on job-satisfaction additional to the effects of teacher self-efficacy and collective-self-efficacy. They do so by applying the Group Actor-Partner Interdependence Model (GAPIM) to cross-sectional questionnaire data from more than 1500 teachers in 37 schools. The GAPIM is an extended form of multilevel analysis. Application of the GAPIM is innovative, because it takes differences in team compositions and the position of individuals within a team into consideration, whereas standard multilevel analysis only takes averaged measures over individuals within teams into consideration. This allows more differentiated analysis of multilevel structures in school improvement research. The findings of this study show that the similarity of an individual teacher to the other teachers in the team, as well as the similarity amongst the other teachers themselves in the team, affects individual teachers' job satisfaction, additional to the effects of self and collective-efficacy.

In Chap. 7, Ng approaches within-group composition from another angle. He conceptualizes schools as social systems and argues that the application of Social Network Analysis is beneficial to understand more about the complexity of

educational leadership. In fact, the author shows that complexity methodologies are neither applied in educational leadership studies, nor are they taught in educational leadership courses. As such, the neglect of complexity methodologies, and there-with also the neglect of innovative insights from the complex and dynamic systems perspective, is reproduced by those who are tasked with, and taught, to empirically assess and foster school improvement. Moreover, the author highlights the mismatch between the assumptions that underlie commonly used inferential statistics and the complexity and dynamics of processes in schools (such as the formation of social ties or adaptation), and describes the resulting problems. Consequently, the author argues for the adoption of complexity methodologies (and dynamic systems tools) and gives an example of the application of Social Network Analysis.

In Chap. 8, Lowenhaupt assesses educational leadership by focusing on the use of language to implement reforms in schools. Applying Rhetorical Analysis (a special case of Discourse Analysis) to data from 14 observations from one case, she undertakes an in-depth investigation of the language of leadership in the implementation of reform. She gives examples of how a school leader's talk could connect more to different audiences' rational, ethical, or affective sides to be more persuasive. The chapter's linguistic turn uncovers aspects of the complexity of school improvement that require more investigation. Moreover, the chapter addresses the importance of sensitivity to one's audience and attuned use of language to foster school improvement.

In Chap. 9, Spillane and Zuberi present yet another methodological innovation to assess educational leadership with: logs. Logs are measurement instruments that can tap into practitioners' activities in a context (and time-point) sensitive manner and can thus be used to understand more about the systematics of (the evolution of) situated micro-processes, such as in this case daily instructional and distributed leadership activities. The specific aim of the chapter is the validation of the Leadership Daily Practice (LDP) log that the authors developed. The LDP log was administered to 34, formal and informal, school leaders for 2 consecutive weeks, in which they were asked to fill in a log-entry every hour. In addition, more than 20 of the participants were observed and interviewed twice. The qualitative data from these three sources were coded and compared. Results from Interrater Reliability Analysis and Frequency Analyses (that were supported by descriptions of exemplary occurrences) suggest that the LDP log validly captures school leaders' daily activities, but also that an extension of the measurement period to encompass an entire school year would be crucial to capture time-point specific variation.

In Chap. 10, Vanblaere and Devos present the use of logs to gain an in-depth understanding of collaboration in teachers' Professional Learning Communities (PLC). Using an explanatory sequential mixed methods design, the authors first administered questionnaires to measure collective responsibility, deprivatized practice, and reflective dialogue and applied Hierarchical Cluster Analysis to the cross-sectional quantitative data from more than 700 teachers in 48 schools to determine the developmental stages of the teachers' PLCs. Based upon the results thereof, 2 low PLC and 2 high PLC cases were selected. Then, logs were administered to the 29 teachers within these cases at four time-points with even intervals over the course

of 1 year. The resulting qualitative data were coded to reflect the type, content, stakeholders, and duration of collaboration. Then, the codes were used in Within and Cross-Case Analyses to assess how the communities of teachers differed in how their learning progressed over time. This study's procedure is a rare example of how the breadth of quantitative research and the depth of qualitative research can thoroughly complement each other to give rich answers to research questions. The findings show that learning outcomes are more diverse in PLCs with higher developmental stages.

In Chap. 11, Oude Groote Beverborg, Wijnants, Slegers, and Feldhoff, use logs to explore routines in teachers' daily reflective learning. This required a conceptualization of reflection as a situated and dynamic process. Moreover, the authors argue that logs do not only function as measurement instruments but also as interventions on reflective processes, and as such might be applied to organize reflective learning in the workplace. A daily and a monthly reflection log were administered to 17 teachers for 5 consecutive months. The monthly log was designed to make new insights explicit, and based on the response rates thereof, an overall insight intensity measure was calculated. This measure was used to assess to whom reflection through logs fitted better and to whom logs fitted worse. The daily log was designed to make encountered environmental information explicit, and the response rates thereof generated dense time-series, which were used in Recurrence Quantification Analysis (RQA). RQA is an analysis techniques with which patterns in temporal variability of dynamic systems can be assessed, such as in this case the stability of the intervals with which each teacher makes information explicit. The innovation of the analysis lies in that it captures how processes of individuals unfold over time and how that may differ between individuals. The findings indicated that reflection through logs fitted about half of the participants, and also that only some participants seemed to benefit from a determined routine in daily reflection.

In Chap. 12, Maag Merki, Grob, Rechsteiner, Wullschleger, Schori, and Rickenbacher applied logs to assess teachers' regulation activities in school improvement processes. First, they developed a theoretical framework based on theories of organizational learning, learning communities, and self-regulated learning. To understand the workings of daily regulation activities, the focus was on how these activities differ between teachers' roles and schools, how they relate to daily perceptions of their benefits and daily satisfaction, and how these relations differ between schools. Second, data about teachers' performance-related, day-to-day activities were gathered using logs as time sampling instruments, a research method that has so far been rarely implemented in school improvement research. The logs were administered 3 times for 7 consecutive days with a 7-day pause between those measurements to 81 teachers. The data were analyzed with Chi-square Tests and Pearson Correlations, as well as with Binary Logistic, Linear, and Random Slope Multilevel Analysis. This study provides a thorough example of how conceptual development, the adoption of a novel measurement instrument, and the application of existing, but elaborate, analyses can be made to interconnect. The results revealed that differences in engagement in regulation activities related to teachers' specific roles, that perceived benefits of regulation activities differed a little between schools, and that those perceived benefits and perceived satisfaction were related.

In Chap. 13, Chaps. 3 through 12 will be discussed in the light of the conceptualization of complexity as presented in Chap. 2. We hope that this book is contributing to the (much) needed specific methodological discourse within school improvement research. We also hope that it will help those who are tasked with empirically assessing and fostering improvements in designing and conducting useful, complex studies on school improvement and accountability.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 2

Why Must Everything Be So Complicated? Demands and Challenges on Methods for Analyzing School Improvement Processes



Tobias Feldhoff and Falk Radisch

2.1 Introduction

In the recent years, awareness has risen by an increasing number of researchers, that we need studies that appropriately model the complexity of school improvement, if we want to reach a better understanding of the relation of different aspects of a school improvement capacity and their effects on teaching and student outcomes, (Feldhoff, Radisch, & Klieme, 2014; Hallinger & Heck, 2011; Sammons, Davis, Day, & Gu, 2014). The complexity of school improvement is determined by many factors (Feldhoff, Radisch, & Bischof, 2016). For example, it can be understood in terms of diverse direct and indirect factors being effective at different levels (e.g., the system, school, classroom, student level), the extent of their reciprocal interdependencies (Fullan, 1985; Hopkins, Ainscow, & West, 1994) and at least the different and widely unknown time periods as well as the various paths school improvement is following in different schools over time to become effective. As a social process, school improvement is also characterized by a lack of standardization and determination (ibid., Weick, 1976). For many aspects that are relevant to school improvement theories, we have only insufficient empirical evidence, especially considering the longitudinal perspective that improvement is going on over time. Valid results depend on plausible theoretical explanations as well as on adequate methodological implementations. Furthermore, many studies could be found to reach contradictory results (e.g. for leadership, see Hallinger & Heck, 1996). In our view, this can at least in part be attributed to the inappropriate consideration of the complexity of school improvement.

T. Feldhoff (✉)
Johannes Gutenberg University, Mainz, Germany
e-mail: feldhoff@uni-mainz.de

F. Radisch
University of Rostock, Rostock, Germany

So far, respective quantitative studies that consider that complexity appropriately have hardly been realized because of the high efforts of current methods and costs involved (Feldhoff et al., 2016). Current elaborate methods, like level-shape, latent difference score (LDS) or multilevel growth models (MGM) (Ferrer & McArdle, 2010; Gottfried, Marcoulides, Gottfried, Oliver, & Guerin, 2007; McArdle, 2009; McArdle & Hamagami, 2001; Raykov & Marcoulides, 2006; Snijders & Bosker, 2003) place high demands on study designs, like large numbers of cases at school-, class- and student-level in combination with more than three well-defined and reasoned measurement points. Not only pragmatic research reasons (benefit-cost-relation, a limit of resources, access to the field) conflict with this challenge. Often, also the field of research cannot fulfil all requirements (for example regarding the needed samples sizes on all levels or the required quantity and intensity of measurement points to observe processes in detail). It is obvious to look for new innovative methods that adequately describe the complexity of school improvement, which at the same time present fewer challenges in the design of the studies. Regarding quantitative research, in the past particularly methods from educational effectiveness research were borrowed. Through this, the complexity of school improvement processes and the resulting demands were not sufficiently taken into account and reflected. Therefore, we need an own methodological and methodical analysis. It is not about inventing new methods but about systematically finding methods in other fields that can adequately handle specific aspects of the overall complexity of school improvement, and that can be combined with other methods that highlight different aspects and, in the end, be able to answer the research questions appropriately. To conduct a meaningful search for new innovative methods, it is first essential to describe the complexity of school improvement and its challenges in detail. This more methodological topic will be discussed in this paper. For that, we present a further development of our framework of the complexity of school improvement (Feldhoff et al., 2016). It helps us to define and to systemize the different aspects of complexity. Based on the framework, research approaches and methods can be systematically evaluated concerning their strong and weak points for specific problems in school improvement. Furthermore, it offers the possibility to search specifically for new approaches and methods as well as to consider even more intensively the combination of different methods regarding their contribution to capturing the complexity of school improvement.

The framework is based upon a systematic long-term review of the school improvement research and various theoretical models that describe the nature of school improvement (see also Fig. 2.1). For that, it might be not settled. As a framework, it shows a wide openness for extending and more differentiating work in the future.

Following this, we will try to draft questions that contribute to classification and critical reflection of new innovative methods, which shall be presented in that book.

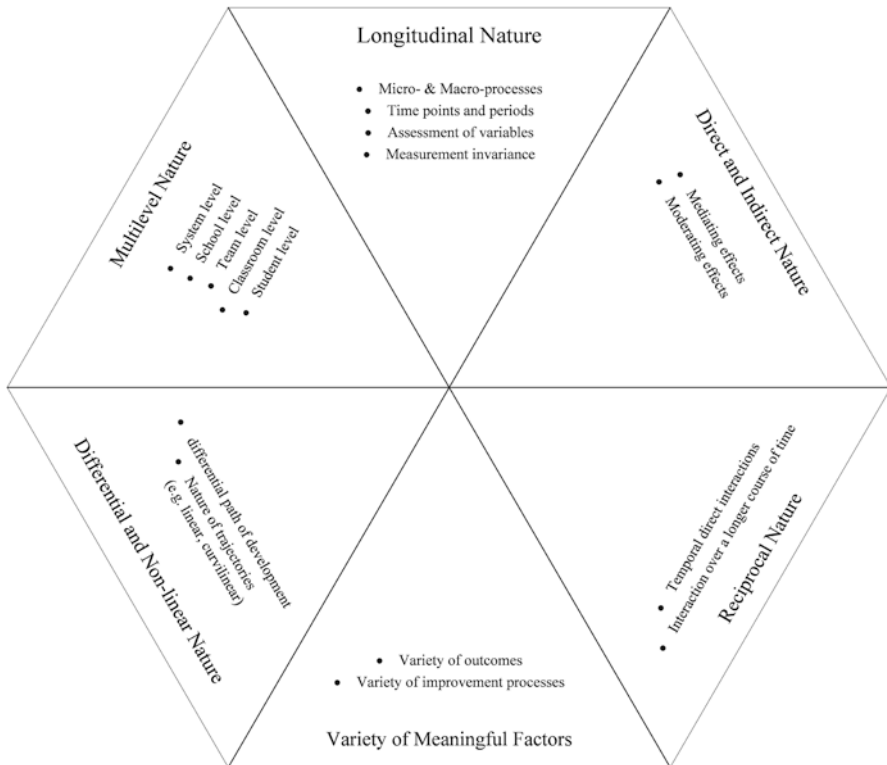


Fig. 2.1 Framework of Complexity

2.2 Theoretical Framework of the Complexity of School Improvement Processes

School improvement targets the school as a whole. As an organizational process, school improvement is aimed at influencing the collective school capacity to change (including change for improvement relevant processes, like cooperation, processes, etc.), the skills of its members, and the students' learning conditions and outcomes (Hopkins, 1996; Maag Merki, 2008; Mulford & Silins, 2009; Murphy, 2013; van Velzen et al., 1985). In order to achieve sustainable school improvement, school practitioners engage in a complex process comprising diverse strategies implemented at the district, school, team and classroom level (Hallinger & Heck, 2011; Mulford & Silins, 2009; Murphy, 2013). School improvement research is interested in both, which processes are involved in which way and what their effects are.

Within our framework the complexity of school improvement as a social process can be described by six characteristics: (a) the longitudinal nature, (b) the indirect nature, (c) the multilevel phenomenon, (d) the reciprocal nature, (e) differential development and nonlinear effects and (f) the variety of meaningful factors (Feldhoff et al., 2016). Explanations of these characteristics are given below:

(a) *The Longitudinal Nature of School Improvement Process*

As Stoll and Fink (1996) pointed out, “Although not all change is improvement, all improvement involves change” (p. 44). Fundamental limitations of the cross-sectional design, therefore, constrain the validity of results when seeking to understand ‘school improvement’ and its related processes. Since school improvement always implies a change in organizational factors (processes and conditions, e.g. behaviours, practices, capacity, attitudes, regulations, and outcomes) over time, it is most appropriately studied in a longitudinal perspective.

It is important to distinguish between changes in micro- and macro-processes. The distinction between micro- and macro-processes is the level of abstraction with which researchers conceptualise and measure practices of actors within schools. Micro-processes are the direct interaction between actors and their practices in the daily work. For example, the cooperation activities of four members of a team in one or more consecutive team meetings. Macro-processes can be described as a sum of direct interactions at a higher level of abstraction and, for the most part, over a longer period of time. For example, what content teachers in a team have exchanged in the last 6 months or about the general way of cooperation in a team (e.g. sharing of materials, joint development of teaching concepts, etc.). While changes of micro processes are possible in a relatively short time, changes of macro-processes can often only be detected and measured after a more extended period (see, e.g. Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; Fullan, 1991; Fullan, Miles, & Taylor, 1980; Smink, 1991; Stoll & Fink, 1996). Stoll and Fink (1996) assume that moderate changes require 3–5 years while more comprehensive changes involve even more extended periods of time (see also Fullan, 1991). The most school improvement studies analyse macro-processes and their effects. But it must also be considered that concrete micro-processes can lead to changes faster due to the dynamical component of interaction and cooperation being more direct and immediate in these processes. Regarding macro-processes, the common way of “aggregation” in micro-processes (usually averaging of quality respectively quantity assessments or their changes) leads to distortions. One phenomenon described adequately in the literature is the one of professional cooperation between teachers. Usually, there are several – parallel – settings of cooperation that can be found in one school. It is highly plausible that already the assessment of the micro-processes in these settings of cooperation turns out to be high-graded different and that this is true in particular for the assessment of changes of micro-processes in these cooperation settings. For example, in individual settings will appear negative changes while in meantime there will be positive changes in others. The usual methods of aggregation to generate characteristics of macro-processes on a higher level are not able to consider these different dynamics – and therefore inevitably lead to distortions.

The rationale for using longitudinal designs in school improvement research is not only grounded in the conceptual argument that change occurs over time (e.g. see Ogawa & Bossert, 1995), but also in the methodological requirements for assigning causal attributions to school policies and practices. Ultimately, school improvement

research is concerned with understanding the nature of relations among different factors that impact on the productive change in desired student outcomes over time (Hallinger & Heck, 2011; Murphy, 2013). The assignment of causal attributions is facilitated by substantial theoretical justification as well as by measurements at different points in time (Finkel, 1995; Hallinger & Heck, 2011; Zimmermann, 1972). “With a longitudinal design, the time ordering of events is often relatively easy to establish, while in cross-sectional designs this is typically impossible” (Gustafsson, 2010, p. 79). Cross-sectional modeling of causal relations might lead to incorrect estimations even if the hypotheses are excellent and reliable. For example, a study investigating the influence of teacher cooperation as macro-processes on student achievement in mathematics demonstrates no effect in cross-sectional analyses, while positive effects emerge in longitudinal modeling (Klieme, 2007).

Recently, Thoonen, Slegers, Oort, and Peetsma (2012) highlighted the lack of longitudinal studies in school improvement research. This lack of longitudinal studies was also observed by Klieme and Steinert (2008) as well as Hallinger and Heck (2011). Feldhoff et al. (2016) have systematically reviewed how common (or rather uncommon) longitudinal studies are in school improvement research. They find only 13 articles that analyzed the relation of school improvement factors and teaching or student outcome longitudinal. Since school improvement research that is based on cross-sectional study designs cannot deliver any reliable information concerning change and its effects on student outcomes, a longitudinal perspective is a central criterion for the power of a study.

Based on the nature of school improvement, the following factors are relevant in longitudinal studies:

Time Points and Period of Development

To investigate a change in school improvement processes and their effects, it is pertinent to consider how often and at which point in time data should be assessed to model the dynamics of the reviewed change appropriately.

The frequency of measurements strongly depends on the different dynamics of change regarding factors. If researchers are interested in the change of micro-processes and their interaction, a higher dynamics is to be expected than those who are interested in changes of macro-processes. A high level of dynamics requires high frequencies (e.g., Reichardt, 2006; Selig et al., 2012). This means that for changes in micro-processes, sometimes daily or weekly measurements with a relatively large number of measurement times (e.g. 10 or 20) are necessary, while for changes of macro-processes, under certain circumstances, significantly less measurement times (e.g. 3–4) suffice, at intervals of several months. Within the limits of research pragmatics, intervals should be accurately determined according to theoretical considerations and previous findings. Furthermore, a critical description and clear justification needs to be given. To identify effects, the period assessed needs to be determined in a way that such effects can be expected from a theoretical point of view (see Stoll & Fink, 1996).

Longitudinal Assessment of Variables

Not only the number of measurement points and the time spans between are relevant for longitudinal studies, but also which of the variables are investigated longitudinally. In many cases, studies often focus solely on a longitudinal investigation of the so-called dependent variable in the form of student outcomes – but concerning conceiving school improvement as change, it is also essential to measure the relevant school improvement factors longitudinally. This is especially important when considering the reciprocal nature of school improvement (see 2.2.4).

Measurement Variance and Invariance

It is highly significant to consider measurement invariance in longitudinal studies (Khoo, West, Wu, & Kwok, 2006, p. 312), because if the meaning of a construct changes, it is empirically not possible to elicit whether change of the construct causes an observed change of measurement scores, change of the reality or an interaction of both (see also Brown, 2006).

For that, the prior testing of the quality of the measuring instruments is more critical and more demanding for longitudinal than for cross-sectional studies. For example, it has to cover the same aspects as well, but in addition with a component that is stable over time. For example, a change of construct-comprehension of the test persons (through learning effects, maturing, etc.) has to be taken into account, and the measuring instruments need to be made robust against these changes for using with common methods. Before the first testing, it is essential to consider which aspects the longitudinal studies should evaluate concerning the improvement processes. Especially more complex school improvement studies present challenges because dynamics can arise and processes gain meaning that cannot be foreseen. That particular dynamic component of the complexity of school improvement can explicitly lead to (maybe intended) changing meanings of constructs by the school improvement processes itself. For example, it is plausible that due to school improvement processes single aspects and items acquiring cooperation between colleagues change concerning their value for participants. Regarding an ideal-typical school improvement process, in the beginning cooperation for a teacher means in particular division of work and exchange of materials and in the end of the process these aspects lost their value and those of joined reflection and preparing lessons as well as trust and a common sense increase. With the help of an according orientation and concrete measures this effect can even be a planned aim of school improvement processes but can also (unwantedly) appear as a side effect of intended dynamical micro-processes. Depending on the involvement and personal interpretation of the gathered experiences, different changes and displacements of attribution of value can be found. – At a moment that will mostly hinder a longitudinal measurement by a lack of measurement invariance across the measurement time points, since most of the methods analysing longitudinal data need a specific measurement invariance.

Many longitudinal studies use instruments and measurement models that were developed for cross-sectional studies (for German studies, this is easily viewable in

the national database of the research data centre (FDZ) Bildung, <https://www.fdz-bildung.de/zugang-erhebungsinstrumente>). Their use is mostly not critically questioned or carefully considered in connection with the specific requirements of longitudinal studies. For psychological research, Khoo, West, Wu and Kwok (2006) recommend more attention to the further consideration of measuring instruments and models. This can be simultaneously transfer to the improvement of measuring instruments for school improvement research.

Measurement invariance touches upon another problem of the longitudinal testing of constructs: The sensitivity of the instruments towards changes that should be observed. The widely used four-level or five-level Likert scales are mostly not sensitive enough towards the different theoretical and empirical expectable developments. They were developed to measure the manifestation or structure of a characteristic on a specific point of time – usually aiming to analyse differences and connections of these manifestation. How and in which dynamic a construct changes over time was not considered in creating Likert scales. For example, cooperation between colleagues, intensity of joined norms and values, the willingness of being innovative are all constructs which are developed out of a cross-sectional perspective in school improvement research. It might be more reasonable to operationalize the construct in a way that can depict various aspects through the course of development, by using the help of different items. Looking at these constructs, for example those of cooperation between colleagues (Gräsel et al., 2006; Steinert et al., 2006) you will often find theoretical deliberations of distinguishing between forms of cooperation and the underlying beliefs. Furthermore, evidences for actual frequency and intensity of cooperation remaining behind their significance are being found again and again not only in the German-speaking field. Concerning school improvement, it is highly plausible that exactly aimed measures can lead to not only increasing amount and intensity of cooperation but also changes in beliefs regarding cooperation which then also lead to a different assessment of cooperation and a displacement of significance of single items and the whole construct itself. It is even assumable that this is the only way of sustainably reaching a substantial increase of intensity and amount of cooperation. A quantitative measure of changes with cross-sectional developed instruments and usual methods is demanding to impossible. We either need instruments, that are stabile in other dimensions to be able of displaying the necessary changes comparably – or methods which are able to portray dynamic construct changes.

(b) *Direct and Indirect Nature of School Improvement*

School improvement can be perceived as a complex process in which changes are initiated at the school level to achieve a positive impact on student learning at the end. It is widely recognized that changes at the school level only become evident after individual teachers have re-contextualized, adapted and implemented them in their classrooms (Hall, 2013; Hopkins, Reynolds, & Gray, 1999; O’Day, 2002). Two aspects of the complexity of school improvement can be deduced from this description, i.e., the direct and indirect nature of school improvement on one hand and the multilevel structure on the other (see 2.2.3).

Depending on the aim, school improvement processes have direct or indirect effects. An example of direct effects is the influence of cooperation on teachers' professionalization. In many respects, school improvement processes involve mediated effects, for instance concerning processes, located in the classroom or even on the team-level that are initiated and/or managed by the school's principal. In school leadership research, Pitner (1988), at an early stage, already stated that the influence of school leadership is indirect and mediated by (1) purposes and goals; (2) structure and social networks; (3) people and (4) organizational culture (Hallinger & Heck, 1998, p. 171). Similar models we can find in school improvement research (Hallinger & Heck, 2011; Slegers et al., 2014). They are based on the assumption that different school improvement factors reciprocally influence each other; some of them directly and others indirectly through different paths (see also: reciprocity). We, moreover, assume that teaching processes are essential mediators of school improvement effects, especially on student outcomes. Ever since school leadership actions have consistently been modeled as mediated effects in school leadership research, a more similar picture of findings has emerged, and a positive impact of school leadership on student outcomes have been found (Hallinger & Heck, 1998; Scheerens, 2012). Also, Hallinger and Heck (1996) and Witziers, Bosker, and Krüger (2003) showed that neglect of mediating factors leads to a lack of validity of the findings, and it remains unclear which effects are being measured. Similar patterns can be expected for the impact of school improvement capacity (see 2.2.6).

(c) School Improvement as a Multilevel Phenomenon

Following Stoll and Fink (1996), we see school improvement as an intentional, planned change process that unfolds at the school level. Its success, however, depends on a change in the actions and attitudes of individual teachers. For example, in the research on professional communities, the actions in teams have a significant impact on those changes (Stoll, Bolam, McMahon, Wallace, & Thomas, 2006). Changes in the actions and attitudes of individual teachers should lead to changes in instruction and the learning conditions of students. These changes should finally have an impact on the students' learning gain. School improvement is a phenomenon that takes place at three or four different known levels within schools (the school level, the team level, the teacher or classroom level, and the student level). It is essential to consider these different levels when investigating school improvement processes (see also Hallinger & Heck, 1998). For school effectiveness research, Scheerens and Bosker (1997, pp. 58) describe various alternative models for cross-level effects, which offer approaches that are also interesting for school improvement research.

Many articles plausibly point out that neither disaggregation at the individual level (that means copying the same number to all members of the aggregate-unit) nor aggregation of information is suitable for taking the hierarchical structure of the data into account appropriately (Heck & Thomas, 2009; Kaplan & Elliott, 1997). School effectiveness research also has widely demonstrated the issues that arise when neglecting single levels. Van den Noortgate, Opdenakker, and Onghena (2005) carried out analyses and simulation studies and concluded that it is essential to not

only take those levels into account where the interesting effects are located. A focus on just those levels might lead to distortions, bearing a negative impact on the validity of the results. Nowadays, multi-level analyses have thus become standard procedure in empirical school (effectiveness) research (Luyten & Sammons, 2010). And it is only a short step postulating that this should become standard in school improvement research too.

Particularly, the combination of micro and macro-processes can only be deduced on methodical ways which adequately display the complex multilevel structure of school (e.g. parallel structures (e.g. classroom vs. team structure), sometimes unclear or instable multilevel structure (e.g. newly initiated or ending team structures every academic year or changings within an academic year), dependent variables on a higher level (e.g. if it is the overall goal to change organisational beliefs), etc.).

(d) *The Reciprocal Nature of School Improvement*

Another aspect, reflecting the complexity of school improvement, evolves from the circumstance that building a school capacity to change and its effects on teaching and student or school improvement outcomes result from reciprocal and interdependent processes. These processes involving different process factors (leadership, professional learning communities, the professionalization of teachers, shared objectives and norms, teaching, student learning) and persons (leadership, teams, teachers, students) (Stoll, 2009). Reciprocity of micro- and macro-processes set differing requirements to the methods (see 2.2.1, longitudinal nature). In micro-processes, there is reciprocity in the way of direct temporal interactions of various persons or factors (within a session/meeting, or days, or weeks). For example, interactions between team members during a meeting enable sense-making and encourage decision-making. In macro-processes, the reciprocity of interactions between various persons or factors is on a more abstract or general level during a longer course of time (maybe several months or years) of improvement processes.

This means, for example, that school leaders not only influence teamwork in professional learning communities over time but also react to changes in teamwork by adapting their leadership actions. Regarding sustainability and the interplay with external reform programs, reciprocity is relevant as a specific form of adaptation to internal and external change. For example, concepts of organizational learning argue that learning is necessary because the continuity and success of organizations depend on their optimal fit to their environment (e.g. March, 1975; Argyris & Schön, 1978). Similar ideas can be found in contingency theory (Mintzberg, 1979) or the context of capacity building for school improvement (Bain, Walker, & Chan, 2011; Stoll, 2009) as well as in school effectiveness research (Creemers & Kyriakides, 2008; Scheerens & Creemers, 1989). School improvement can thus be described as a process of adapting to internal and external conditions (Bain et al., 2011; Stoll, 2009). The success of schools and their improvement capacity is thus a result of this process.

The empirical investigation of reciprocity requires designs that assess all relevant factors of the school improvement process, mediating factors (for example instructional factors) and outcomes (e.g. student outcomes) at several measurement points, in a manner that allows to model effects in more than one direction.

(e) *Differential Paths of Development and Nonlinear Trajectories*

The fact that the development of an improvement capacity can progress in very different ways adds to the complexity of school reform processes (Hopkins et al., 1994; Stoll & Fink, 1996). Because of their different conditions and cultures, schools differ in their initial levels, the strength and in the progress of their development. The strength and progress of the development itself depends also from the initial level (Hallinger & Heck, 2011). In some schools, development is continuous while in other cases an implementation dip is observable (e.g., Bryk et al., 2010; Fullan, 1991). Theoretically, many developmental trajectories are possible across time, many of which are presumably not linear.

Nonlinearity does not only affect the developmental trajectories of schools. It can be assumed that many relationships between school improvement processes among themselves or in relation to teaching processes and outcomes are also not linear (Creemers & Kyriakides, 2008). Often curvilinear relationships can be expected, in which there is a positive relation between two factors up to a certain point. If this point is exceeded, the relation is near zero or zero, or it can become negative. The first case, the relation becomes zero or near zero, can be interpreted as a kind of a saturation effect. For example, theoretically, it can be assumed that the willingness to innovate in a school, at a certain level, has little or no effect on the level of cooperation in the school. An example of a positive relationship that becomes negative at some level is the correlation between the frequency and intensity of feedback and evaluation on the professionalization of teachers. In the case of a successful implementation, it can be assumed that the frequency and intensity of feedback and evaluation will have a positive effect on the professionalization of teachers. If the frequency and intensity exceed a certain level, it can be assumed that the teachers feel controlled, and the effort involved with the feedback exceeds the benefits and thus has a negative effect on their professionalization. Where the level is set which is critical for each individual school and when it is reached, is dependent on the interaction of different factors on the level of micro- and macro-processes (teachers feeling assured, frustration tolerance, type and acceptance of the style of leadership, etc.). With this example, it also gets clear that there does not only exist no “the more the better” in our concept but also the type and grade of an “ideal level” is dependent on the dynamical and reciprocal interaction with other factors in the duration of time and on the context of the considered actors. Currently, our understanding of the nature of many relationships of school improvement processes among themselves, or in relation to teaching and outcomes is very low (Creemers & Kyriakides, 2008).

To map this complexity, methods are required that enable modelling of nonlinear effects as well as individual development. In empirical studies, it is necessary to examine the course of developments and correlations – whether they are linear,

curvilinear or better describable and explainable via sections or threshold phenomena (e.g. by comparing different adaptation functions in regressive evaluation methods, sequential analysis of extensive longitudinal sections a variety of measurements, etc.). Particularly valuable are procedures that justify several alternative models in advance and test them against each other. Such approaches could improve understanding of changes in school improvement research. But however, these methods (e.g. nonlinear regressive models) have never been used in school improvement research nor in school effective research. The same applies to the study of the variability of processes, developments and contexts. Particularly in recent years, for example, with growth curve models and various methods of multi-level longitudinal analysis, numerous new possibilities have been established in order to carry out such investigations. They also open up the possibility of looking at and examining the variability of processes as dependent variables.

The analysis of different development trajectories of schools and how these correlate e.g. with the initial level and the result of the development is obviously highly relevant for educational accountability and the evaluation of reform projects. In many pilot projects or reform programs, too little consideration is given to the different initial levels of schools and their developmental trajectories. This often leads to Matthew effects. However, reforms in their implementation can only take those factors into account, if the appropriate knowledge about them has been generated in advance in corresponding evaluation studies.

(f) *Variety of Meaningful Factors*

Many different persons and processes are involved in changes in school and their effects on student outcomes (e.g., Fullan, 1991; Hopkins et al., 1994; Stoll, 2009 see 2.2.2). The diversity of factors relates to all parts of the process (e.g. improvement capacity, student outcomes, and teaching, contexts). Because this chapter (and this book) deals with school improvement, we want to confine ourselves exemplarily to two central parts. On one hand, we focus on the variety of factors of improvement processes, because we want to show that in this central part of school improvement a reduction of the variety of factors is not easily achieved. On the other hand, we focus on the variety of outcomes/outputs, since we want to contribute to the still emerging discussion about a stronger merging of school improvement research and school effectiveness research.

Variety of Factors of Improvement Capacity

As outlined above, school improvement processes are social processes that cannot be determined in a clear-cut way. School improvement processes are diverse and interdependent, and they might involve many aspects in different ways. It is essential to consider the variety and reciprocity of meaningful factors of a school's improvement capacity (e.g., teacher cooperation, shared meaning and values, leadership, feedback, etc.) while investigating the relation of different school improvement aspects and their outcomes. A neglect of this diversity can lead to a false estimation of the effects. Only by considering all meaningful factors of the

improvement capacity, it will be possible to take into account interactions between the factors as well as shared, interdependent and/or possibly contradictory effects.

By merely looking at individual aspects, researchers might fail to identify effects that only result from interdependence. Another possible consequence might be a mistaken estimation of factors.

Variety of Outcomes

Given the functions, schools hold for society and the individual, a range of school-related outputs and outcomes can be deduced. The effectiveness of school improvement has been left unattended for a long time. Different authors and sources claim that school effectiveness research and school improvement research should cross-fertilize (Creemers & Kyriakides, 2008). One of the central demands is to make school improvement research more effective in a way that includes all societal and individual spheres of action. Under such a broad perspective that is necessarily connected with school improvement research, it is clear, that focusing on student-related outcomes (what itself means more than cognitive outcomes) is only exemplary (Feldhoff, Bischof, Emmerich & Radisch, 2015; Reezigt & Creemers, 2005). Scheerens and Bosker (1997) distinguish short-term outputs and long-term outcomes (pp. 4). Short-term outputs comprise cognitive as well as motivational-affective, metacognitive and behavioural criteria (Seidel, 2008). The diversity of short-term outputs suggests that the different aspects of the capacity are correlated in different ways to individual output criteria via different paths. Findings on the relation of capacity to one output cannot automatically be transferred to other output aspects or outcomes. If we wish to understand school improvement, we need to consider different aspects of school output in our studies. Seidel (2008) has demonstrated that school effectiveness research at the school level is almost exclusively limited to cognitive subject-related learning outcomes (see also Reynolds, Sammons, De Fraine, Townsend, & Van Damme, 2011). Seidel indicates that the call for consideration of multi-criterial outcomes in school effectiveness research has hardly been addressed (see p. 359). In this regard, so far little if anything is known about the situation in school improvement research.

2.3 Conclusion and Outlook

The framework systematically shows the complexity of school improvement processes in its six characteristics and which methodological aspects need to be considered when developing a research design and choosing methods. Like we drafted in the introduction it is for example due to limited resources and limited access to schools not always possible to consider all aspects similarly. Nevertheless, it is important to reflect and reason: Which aspects can not or only limited be considered, what effects emerge on knowledge acquisition and the results out of this non-consideration or limited consideration of aspects and why a limited or non-consideration is despite limits in terms of knowledge acquisition still reasonable.

In this sense, unreflect or inadequate simplification and thus inappropriate modelling might lead to empirical results and theories that do not face reality or that are leading to contradictory findings. In sum it will lead to a stagnation in the further development of theoretical models. A reliable and further development would require the recognition and the exclusion of inappropriate consideration of the complexity as a cause of contradictory findings. Our methods and designs influence our perspectives as they are the tools by which we generate knowledge, which in turn is the basis for constructing, testing and enhancing our theoretical models (Feldhoff et al., 2016).

Therefore, it is time to search for new methods that make it possible to consider the aspects of complexity, and that has not been made use of in the research of school improvement so far. Many quantitative and qualitative methods have emerged over the last decades within various disciplines of social sciences that need to be reflected for their usefulness and practicality for the school improvement research. To ease the systematic search for adequate and useful methods, we formulated questions based on the theoretical framework, that helps to review the methods' usefulness overall critically and for every single aspect of the complexity. They can also be used as guiding questions for the following chapters.

2.3.1 Guiding Questions

Longitudinal

- Can the method handle longitudinal data?
- Is the method more suitable for shorter or longer intervals (periods between measurement points)?
- How many measurement points are needed and how many are possible to handle?
- Is it affordable to have similar measurement points (comparable periods between the single measurement points and same measurement points for all the individuals and schools)?
- Is the method able to measure all variables of interest in a longitudinal way?
- Is the method able to differentiate the reasons for (in-)variant measurement scores over time, or does the method handle the possible reasons for the (in-) variation of measurements in any other useful way?

Indirect Nature of School Improvement

- Is the method able to evaluate different ways of modeling indirect paths/effects (e.g., mediation, moderation in one or more steps)?
- Is the method able to measure different paths (direct and indirect) between different variables at the same time?

Multilevel

- Is the method able to handle all the different needed levels of school improvement?
- Is the method able to consider effects at levels that are not of interest?

- Is the method able to consider multilevel effects from a lower level of the hierarchy to a higher level?
- Is the method able to handle more complex structures of the sample (e.g., single or maybe multiple-cross and/or multi-group-classified data)?

Reciprocity

- Is the method able to model reciprocal or other non-single-directed paths?
- Is the method able to model circular paths with unclear differentiating between dependent and independent variables?
- Is the method able to analyze effects on both side of a reciprocal relation at the same time?

Differential Paths and Nonlinear Effects

- Is the method able to handle different paths over time and units?
- What kind of effects can the method handle at the same time (linear, non-linear, different positioning over the time points at different units)?

Variety of Factors

- Is the method able to handle a variety of independent factors with different meanings on different levels?
- Is the method able to handle different dependent factors and does not only focus on cognitive or on another measurable factor at the student level?

In addition to these questions on the individual aspects of complexity, it is also essential to consider to what extent the methods are also suitable for capturing several aspects. Alternatively, with which other methods the method can be combined to take different aspects into account.

Overall Questions

Strengths, Weaknesses, and Innovative Potential

- In which aspects of the complexity of school improvement are the strengths and weaknesses of the method (general and comparable to established methods)?
- Does the method offer the potential to map one or more aspects of the complexity in a way that was previously impossible with any of the “established” methods?
- Is the method more suitable for generating or further developing theories, or rather for checking existing ones?
- What demands does the method make on the theories?
- With which other methods can the respective method be well combined?

Requirements/Cost-Benefit-Ratio

- Which requirements (e.g., numbers of cases at school-, class- and student-level, amount and rigorous timing of measurement points, data collection) put the methods to the design of a study?
- Are the requirements realistic to implement such a design (e.g., concerning finding, enough schools/realizing data collection, get funding)?
- What is the cost-benefit-ratio compared to established methods?

References

- Argyris, C., & Schön, D. (1978). *Organizational learning – A theory of action perspective*. Reading, MA: Addison-Wesley.
- Bain, A., Walker, A., & Chan, A. (2011). Self-organisation and capacity building: Sustaining the change. *Journal of Educational Administration*, 49(6), 701–719.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research: Methodology in the social sciences*. New York, NY: Guilford Press.
- Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.
- Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London, UK/New York, NY: Routledge.
- Feldhoff, T., Bischof, L. M., Emmerich, M., & Radisch, F. (2015). Was nicht passt, wird passend gemacht! Zur Verbindung von Schuleffektivität und Schulentwicklung. In H. J. Abs, T. Brüsemeister, M. Schemmann, & J. Wissinger (Eds.), *Governance im Bildungswesen – Analysen zur Mehrebenenperspektive, Steuerung und Koordination* (pp. 65–87). Wiesbaden, Germany: Springer.
- Feldhoff, T., Radisch, F., & Bischof, L. M. (2016). Designs and methods in school improvement research. A systematic review. *Journal of Educational Administration*, 2(54), 209–240.
- Feldhoff, T., Radisch, F., & Klieme, E. (2014). Methods in longitudinal school improvement research: State of the art. *Journal of Educational Administration*, 52(5), 565–736.
- Ferrer, E., & McArdle, J. J. (2010). Longitudinal Modeling of developmental changes in psychological research. *Current Directions in Psychological Science*, 19(3), 149–154.
- Finkel, S. E. (1995). *Causal analysis with panel data*. Thousand Oaks, CA: Sage.
- Fullan, M. G. (1985). Change processes and strategies at the local level. *The Elementary School Journal*, 85(3), 391–421.
- Fullan, M. G. (1991). *The new meaning of educational change*. New York, NY: Teachers College Press.
- Fullan, M. G., Miles, M. B., & Taylor, G. (1980). Organization development in schools: The state of the art. *Review of Educational Research*, 50(1), 121–183.
- Gottfried, A. E., Marcoulides, G. A., Gottfried, A. W., Oliver, P. H., & Guerin, D. W. (2007). Multivariate latent change modeling of developmental decline in academic intrinsic math. *International Journal of Behavioral Development*, 31(4), 317–327.
- Gräsel, C., Fussangel, K., & Parchmann, I. (2006). Lerngemeinschaften in der Lehrerfortbildung. *Zeitschrift für Erziehungswissenschaft*, 9(4), 545–561.
- Gustafsson, J. E. (2010). Longitudinal designs. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 77–101). Abingdon, UK/New York, NY: Routledge.
- Hall, G. E. (2013). Evaluating change processes: Assessing extent of implementation (constructs, methods and implications). *Journal of Educational Administration*, 51(3), 264–289.
- Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980–1995. *Educational Administration Quarterly*, 32(1), 5–44.
- Hallinger, P., & Heck, R. H. (1998). Exploring the principal's contribution to school effectiveness: 1980–1995. *School Effectiveness and School Improvement*, 9(2), 157–191.
- Hallinger, P., & Heck, R. H. (2011). Conceptual and methodological issues in studying school leadership effects as a reciprocal process. *School Effectiveness and School Improvement*, 22(2), 149–173.
- Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques (Quantitative methodology series)* (2nd ed.). New York: Routledge.
- Hopkins, D. (1996). Towards a theory for school improvement. In J. Gray, D. Reynolds, C. T. Fitz-Gibbon, & D. Jesson (Eds.), *Merging traditions. The future of research on school effectiveness and school improvement* (pp. 30–50). London, UK: Cassell.

- Hopkins, D., Ainscow, M., & West, M. (1994). *School improvement in an era of change*. London, UK: Cassell.
- Hopkins, D., Reynolds, D., & Gray, J. (1999). Moving on and moving up: Confronting the complexities of school improvement in the improving schools project. *Educational Research and Evaluation*, 5(1), 22–40.
- Kaplan, D., & Elliott, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(1), 1–24.
- Khoo, S.-T., West, S. G., Wu, W., & Kwok, O.-M. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement* (pp. 301–317). Washington, DC: American Psychological Association.
- Klieme, E. (2007). *Von der "Output-" zur "Prozessorientierung": Welche Daten brauchen Schulen zur professionellen Entwicklung ?* Kassel: GEPF-Jahrestagung.
- Klieme, E., & Steinert, B. (2008). Schulentwicklung im Längsschnitt. Ein Forschungsprogramm und erste explorative analysen. In M. Prenzel & J. Baumert (Eds.), *Vertiefende Analysen zu PISA 2006* (pp. 221–238). Wiesbaden, Germany: VS Verlag.
- Luyten, H., & Sammons, P. (2010). Multilevel modelling. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 246–276). Abingdon, UK/New York, NY: Routledge.
- Maag Merki, K. (2008). Die Architektur einer Theorie der Schulentwicklung – Voraussetzungen und Strukturen. *Journal für Schulentwicklung*, 12(2), 22–30.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60, 577–605.
- McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 139–175). Washington, DC: American Psychological Association.
- Mintzberg, H. (1979). *The structuring of organizations: A synthesis of the research*. Englewood Cliffs, NJ: Prentice-Hall.
- Mulford, B., & Silins, H. (2009). Revised models and conceptualization of successful school principalship in Tasmania. In B. Mulford & B. Edmunds (Eds.), *Successful school principalship in Tasmania* (pp. 157–183). Launceston, Australia: Faculty of Education, University of Tasmania.
- Murphy, J. (2013). The architecture of school improvement. *Journal of Educational Administration*, 51(3), 252–263.
- O'Day, J. A. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, 72(3), 293–329.
- Ogawa, R. T., & Bossert, S. T. (1995). Leadership as an organizational quality. *Educational Administration Quarterly*, 31(2), 224–243.
- Pitner, N. J. (1988). The study of administrator effects and effectiveness. In N. J. Boyan (Ed.), *Handbook of research on educational* (pp. S. 99–S.122).
- Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Reezigt, G. J., & Creemers, B. P. M. (2005). A comprehensive framework for effective school improvement. *School Effectiveness and School Improvement*, 16(4), 407–424.
- Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods*, 11(1), 1–18.
- Reynolds, D., Sammons, P., De Fraine, B., Townsend, T., & Van Damme, J. (2011). *Educational effectiveness research. State of the art review*. Paper presented at the annual meeting of the international congress for school effectiveness and improvement, Cyprus, January.
- Sammons, P., Davis, S., Day, C., & Gu, Q. (2014). Using mixed methods to investigate school improvement and the role of leadership. *Journal of Educational Administration*, 52(5), 565–589.
- Scheerens, J. (Ed.). (2012). *School leadership effects revisited. Review and meta-analysis of empirical studies*. Dordrecht, Netherlands: Springer.

- Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.
- Scheerens, J., & Creemers, B. P. M. (1989). Conceptualizing school effectiveness. *International Journal of Educational Research*, 13(7), 691–706.
- Seidel, T. (2008). Stichwort: Schuleffektivitätskriterien in der internationalen empirischen Forschung. *Zeitschrift für Erziehungswissenschaft*, 11(3), 348–367.
- Selig, J. P., Preacher, K. J., & Little, T. D. (2012). Modeling time-dependent association in longitudinal data: A lag as moderator approach. *Multivariate Behavioral Research*, 47(5), 697–716. <https://doi.org/10.1080/00273171.2012.715557>
- Slegers, P. J., Thoonen, E., Oort, F. J., & Peetsma, T. D. (2014). Changing classroom practices: The role of school-wide capacity for sustainable improvement. *Journal of Educational Administration*, 52(5), 617–652. <https://doi.org/10.1108/JEA-11-2013-0126>
- Smink, G. (1991). The Cardiff congress, ICSEI 1991. *Network News International*, 1(3), 2–6.
- Snijders, T., & Bosker, R. J. (2003). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.
- Steinert, B., Klieme, E., Maag Merki, K., Döbrich, P., Halbheer, U., & Kunz, A. (2006). Lehrkooperation in der Schule: Konzeption, Erfassung, Ergebnisse. *Zeitschrift für Pädagogik*, 52(2), 185–204.
- Stoll, L. (2009). Capacity building for school improvement or creating capacity for learning? A changing landscape. *Journal of Educational Change*, 10(2–3), 115–127.
- Stoll, L., Bolam, R., McMahon, A., Wallace, M., & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of Educational Change*, 7(4), 221–258.
- Stoll, L., & Fink, D. (1996). *Changing our schools. Linking school effectiveness and school improvement*. Buckingham, UK: Open University Press.
- Thoonen, E. E., Slegers, P. J., Oort, F. J., & Peetsma, T. T. (2012). Building school-wide capacity-for improvement: The role of leadership, school organizational conditions, and teacher factors. *School Effectiveness and School Improvement*, 23(4), 441–460.
- Van den Noortgate, W., Opdenakker, M.-C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, 16(3), 281–303.
- van Velzen, W. G., Miles, M. B., Ekholm, M., Hameyer, U., & Robin, D. (1985). *Making school improvement work: A conceptual guide to practice*. Leuven and Amersfoort: Acco.
- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 21(1), 1–19.
- Witziers, B., Bosker, R. J., & Krüger, M. L. (2003). Educational leadership and student achievement: The elusive search for an association. *Educational Administration Quarterly*, 39(3), 398–425.
- Zimmermann, E. (1972). *Das Experiment in den Sozialwissenschaften*. Stuttgart, Germany: Poeschel.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 3

School Improvement Capacity – A Review and a Reconceptualization from the Perspectives of Educational Effectiveness and Educational Policy



David Reynolds and Annemarie Neeleman

3.1 Introduction

The field of school improvement (SI) has developed rapidly over the last 30 years, moving from the initial Organisational Development (OD) tradition to school-based review, action research models, and the more recent commitment to leadership-generated improvement by way of instructional (currently) and distributed (historically) varieties. However, it has become clear from the findings in the field of educational effectiveness (EE) (Chapman et al., 2012; Reynolds et al., 2014) that SI needs to be aware of the following developmental needs based on insights from both EE (Chapman et al., 2012; Reynolds et al., 2014) and educational practice as well as other research disciplines, if it will be considered an agenda-setting topic for practitioners and educational systems.

3.1.1 What Kind of School Improvement?

Following Scheerens (2016), we interpret school improvement as the “dynamic application of research results” that should follow the research activity of educational effectiveness. Basically, it is the schools and educational systems that have been carrying out school improvement themselves over the years. However, this is poorly understood, rarely conceptualised/measured and, what is even more remarkable, seldom used as the design foundations of conventionally described SI. Many

D. Reynolds (✉)
Swansea University, Swansea, UK
e-mail: david@davidreynoldsconsulting.com; david.reynolds@swansea.ac.uk

A. Neeleman
Maastricht University, Maastricht, The Netherlands

policy-makers and educational researchers tend to cling to the assumption that EE, supported by statistically endorsed effectiveness-enhancing factors, should set the SI agenda (e.g. Creemers & Kyriakides, 2009). However logical this assumption may sound, educational practice has not been necessarily predisposed to act accordingly.

A recent comparison (Neeleman, 2019a) between effectiveness-enhancing factors from three effectiveness syntheses (Hattie, 2009; Robinson, Hohepa, & Lloyd, 2009; Scheerens, 2016) and a data set of 595 school interventions in Dutch secondary schools (Neeleman, 2019b) shows a meagre overlap between certain policy domains that are present in educational practice - especially in organisational and staff domains - and those interventions currently focussed on in EE research. Vice versa, there are research objects in EE that hardly make it to educational practice, even those with considerable effect sizes, such as self-report grades, formative evaluation, or problem-solving teaching.

How are we to interpret and remedy this incongruity? We know from previous research that educational practice is not always predominantly driven by the need to have an increase in school and student outcomes as measured in cognitive tests (often maths and languages) - the main effect size of most EE. We are also familiar with the much-discussed gap between educational research and educational practice (Broekkamp & Van Hout-Wolters, 2007; Brown & Greany, 2017; Levin, 2004; Vanderlinde & Van Braak, 2009) – two clashing worlds speaking different languages and with only few interpreters around. In this paper, we argue for a number of changes in SI to enhance its potential for improving students' chances in life. These changes in SI refer to the context (2), the classroom and teaching (3), the development of SI capacity (4), the interaction with communities (5), and the transfer of SI research into practice (6).

3.2 Contextually Variable School Improvement

Throughout their development, SI and EE have had very little to say about whether or not 'what works' is different in different educational contexts. This happened in part since the early EE discipline had an avowed 'equity' or 'social justice' commitment. This led to an almost exclusive focus in research in many countries on the schools that disadvantaged students attended, leading to the absence of school contexts of other students being in the sampling frame. At a later time, this situation has changed, with most studies now being based upon more nationally representative samples, and with studies attempting to focus on establishing 'what works' across these broader contexts (Scheerens, 2016).

Looking at EE, we cannot emphasize enough that many findings are based on studies conducted in primary education in English-speaking and highly developed countries - mostly, but not exclusively, in the US (Hattie, 2009). From Scheerens (2016, p. 183), we know that "positive findings are mostly found in studies carried out in the United States." Nevertheless, many of the statistical relationships

established in EE over time between school characteristics and student outcomes are on the low side in most of the meta analyses (e.g. Hattie, 2009; Marzano, 2003) with a low variance in outcomes being explained by the use of single school-level factors or averaged groups of them overall.

Strangely, this has insufficiently led to what one might have expected – the disaggregation of samples into smaller groups of schools in accordance with characteristics of their contexts, like socioeconomic background, ethnic (or immigrant) background, urban or rural status, and region. With disaggregation and analysis by groups of schools within these different contexts, it is possible that there could be better school-outcome relationships than overall exist across all contexts with school effects seen as moderated by school context.

This point is nicely made by May, Huff, and Goldring (2012) in an EE study that failed to establish strong links between principals' behaviours and attributes in terms of relating the time spent by principals on various activities and student achievement over time leading to the authors' conclusion that "...contextual factors not only have strong influences on student achievement but also exert strong influences on what actions principals need to take to successfully improve teaching and learning in their schools" (p. 435).

The authors rightly conclude in a memorable paragraph that,

...our statistical models are designed to detect only systemic relationships that appear consistently across the full sample of students and schools. [...] if the success of a principal requires a unique approach to leadership given a school's specific context, then simple comparisons of time spent on activities will not reveal leadership effects on student performance. (also p. 435)

3.2.1 The Role of Context in EE over the Last Decades

In the United States, there was an historic focus on simple contextual effects. Their early definition thereof as 'group effects' on educational outcomes was supplemented in the 1980s and 1990s by a focus on whether the context of the 'catchment area' of the school influenced the nature of the educational factors that schools used to increase their effectiveness. Hallinger and Murphy's (1986) study of 'effective' schools in California, which pursued policies of active parental disinvolvement to buffer their children from the influences of their disadvantaged parents/caregivers, is just one example of this focus. The same goes for the Louisiana School Effectiveness Study (LSES) of Teddlie and Stringfield (1993). Furthermore, there has also been an emphasis in the UK upon how schools in low SES communities need specific policies, such as the creation of an orderly structured atmosphere in schools, so that learning can take place (see reviews in Muijs, Harris, Chapman, Stoll, & Russ, 2004; Reynolds et al., 2014). Also, in the UK, the 'site' of ineffective schools was the subject of intense speculations for a while within the school improvement community in terms of different, specific interventions that were needed due to their distinctive pathology (Reynolds, 2010; Stoll & Myers, 1998).

However, this flowering of what has been called a ‘contingency’ perspective did not last very long. The initial *International Handbook of School Effectiveness Research* (Teddle & Reynolds, 2000) comprises a substantial chapter on ‘context specificity’ whereas the 2016 version does not (Chapman et al., 2016).

Subsequently, many of the lists that were compiled in the 1990s concerning effective school factors and processes had been produced using research grants from official agencies that were anxious to extract ‘what works’ from the early international literature on school effectiveness in order to directly influence school practices. In that context, researchers recognised that acknowledging the findings from schools that showed different process factors being effective in different ways in different contextual areas, would not give the funding bodies what they wanted. Many of the lists were designed for practitioners, who might appreciate the universal mechanisms about ‘what works.’ There was a tendency to report confirmatory findings rather than disconfirmatory ones, which could have been considered ‘inconvenient.’ The school effectiveness field wanted to show that it had alighted on truths: ‘well, it all depends upon context’ was not a view that we believed would be respected by policy and practice. The early EE tradition that showed that ‘what works’ was different in different contexts had largely vanished.

Additional factors reinforced the exclusion of context in the 2000s. First, the desire to ape the methods employed within the much-lauded medical research community – such as experimentation and RCTs – reflected a desire, as in medicine, to be able to intervene in all educational settings with the same, universally applicable methods (as with a universal drug for all illness settings, if one were to exist). The desire to be effective across all school contexts – ‘wherever and whenever we choose’ (Edmonds, 1979, cited in Slavin, 1996) – was a desire for universal mechanisms. Yet, of course, the medical model of research is in fact designed to generate universally powerful interventions and, at the same time, is committed to context specificity with effective interventions being tailored to the individual patient’s context in terms of the kind of drug used (for example one of the forty variants of statin), dosage of the drug, length of usage of the drug, combination of a drug with other drugs, the sequence of usage if combined with other drugs, and patient-dependent variables, like gender, weight, and age. We did not understand this in EE – or perhaps we did comprehend this, but this was not a convenient stance for our future research designs and funding. We picked up on the ‘universal’ applicability but not on the contextual variations. Perhaps we also did not sufficiently recognise the major methodological issues about randomised controlled trials themselves – particularly the issues that deal with sample atypicality.

Second, the meta-analyses that were undertaken ignored contextual factors in the interests of substantial effect sizes. Indeed, national context and local school SES context were rarely factors used to split the overall sample sizes, and (when they did) were based upon superficial operationalization of context (e.g. Scheerens, 2016).

Third, the rash of internationally based studies that attempted to look for regularities cross-culturally in the characteristics of effective schools, and school systems were also of the ‘one right way’ variety. The operationalization of what were

usually highly abstract formulations – such as a ‘guiding coalition’ or group of influential educational persons in a society – was never sufficiently detailed to permit testing of ideas.

Fourth, the run-of-the-mill multilevel, multivariate EE studies analysing whole samples did not disaggregate into SES contexts, urban/rural contexts, or ethnic (or immigrant) background as this would have cut the sample size. Hence, context was something that – as a field – we controlled out in our analyses, not something that we kept in in order to generate more sensitive, multi-layered explanations.

Finally, many of the nationally based educational interventions generated within many Anglo-Saxon societies that were clearly informed by the EE literature involved intervening in disadvantaged, low-SES communities, but with programmes derived from studies that had researched and analysed their data for all contexts, universally. The circle was complete from the 1980s and 1990s research: Specific contexts received programmes generated from universally based research.

It is possible that for understandable reasons, a tradition in educational effectiveness that would have been involved in studying the complex interaction between context and educational processes, and that would have also generated further knowledge about ‘what works by context’, has eroded. This tradition needs to be rebuilt and placed in many educational contexts and applied in school improvement.

3.2.2 Meaningful Context Variables for SI

What contextual factors might provide a focus for a more ‘contingently orientated’ SI approach to ‘what works’ to improve schools? The socio-economic composition of the ‘catchment areas’ of schools is just one important contextual variable – others are whether schools are urban or rural or ‘mixed,’ the level of effectiveness of the school, the trajectory of improvement (or decline) in school results over time, and the proportion of students from a different ethnic (or immigrant) background. Various of these areas have been explored – by Hallinger and Murphy (1986), Teddlie and Stringfield (1993), and Muijs et al. (2004) on SES contextual effects, and by Hopkins (2007), for example, in terms of the effects of where an individual school may be within its own performance cycle affecting what needs to be done to improve.

Other contextual factors that may indicate a need for different interventions in what is needed to improve include:

- Whether the school is primary or secondary for the student age groups covered and/or whether the school is of a distinct organizational type (e.g. selective);
- Whether the school is a member of educational improvement networks;
- Whether the school has significant within-school variation in outcomes, such as achievement that may act as a brake upon any improvement journey, or which could, contrastingly, provide a ‘benchmarking’ opportunity.

- Other possible factors concerning cultural context are:
 - school leadership
 - teacher professionalism/culture
 - complexity of student population (other than SES; regarding inclusive education) and that of parents
 - financial position
 - level of school autonomy and market choice mechanisms
 - position within larger school board/academy and district level “quality” factors

We must conclude by saying that for SI, we simply do not know the power of contextually variable SI.

3.3 School Improvement and Classrooms/Teaching

The importance of the classroom level by comparison with that of the school has so far not been marked by the volume of research that is needed in this area. In all multilevel analyses undertaken, the amount of variance explained by classrooms is much greater than that of the school (see for example Muijs & Reynolds, 2011); yet, it is schools that have generally received more attention from researchers in both SI and EE.

Research into classrooms poses particular problems for researchers. Observation of teachers’ teaching is clearly essential to relate to student achievement scores, but in many societies access to classrooms may be difficult. Observation is time-consuming, as it is important (ethically) to involve briefing and debriefing of research (methods) to individual teachers and parents. The number of instruments to measure teaching has been limited, with the early American instruments of the ‘process-product’ tradition being supplemented by a limited number of instruments from the United Kingdom (e.g. Galton, 1987; Muijs & Reynolds, 2011) and from international surveys (Reynolds, Creemers, Stringfield, Teddlie, & Schaffer, 2002). The insights of PISA studies, and, of course, those of the International Association for the Evaluation of Educational Achievement (IEA), such as TIMMS and PIRLS, say very little about teaching practices because they measure very little about them, with the exception of TALIS.

Instructional improvement at the level of the teacher/teaching is relatively rare, although there have been some ‘instructionally based’ efforts, like those of Slavin (1996) and some of the experimental studies that were part of the old ‘process-product’ tradition of teacher effectiveness research in the United States in the 1980s and 1990s.

However, it seems that SI researchers and practitioners are content to pull levers of intervention that operate mostly at the school level, even though EE repeatedly has shown that they will have less effect than classroom or classroom/school-based ones. It should be mentioned that the problems of adopting a school-based rather than a classroom-based approach have been magnified by the use of multilevel

modelling from the 1990s onwards, which only allocates variance ‘directly’ to different levels rather than looking at the variance explained by the interaction between levels (of school and classroom potentiating each other).

3.3.1 Reasons for Improving Teaching to Foster SI

Research in teaching and the improvement of pedagogy are also needed in order to deal with the further implications of the rapidly growing field of cognitive neuroscience, which has been generated by brain imaging technology, such as Magnetic Resonance Imaging (MRI). Interestingly, the field of cognitive neuroscience has been generated by a methodological advance in just the same way that EE was generated by one, in this latter case, value-added analyses.

Interesting evidence from cognitive neuroscience includes:

- Spaced learning, with suggestions that use of time spaces in lessons, with or without distractor activities, may optimise achievement;
- The importance of working or short-term memory not being overloaded, thereby restricting capacity to transfer newly learned knowledge/skills to long-term memory;
- The evidence that a number of short learning sessions will generate greater acquisition of capacities than more rare, longer sessions -the argument for so-called ‘distributed practice’;
- The relation between sleep and school performance in adolescents (Boschloo et al., 2013).

So, given the likelihood of the impact of neuroscience being major in the next decade, it is the classroom that needs to be a focus as well as the school ‘level’. School improvement, historically, even in its recent manifestation, has been poorly linked – conceptually and practically – with the classroom or ‘learning level’.

The great majority of the improvement ‘levers’ that have been pulled historically are all at the school level, such as through development planning or whole school improvement planning, and although there is a clear intention in most of these initiatives for classroom teaching and student learning to be impacted upon, the links between the school level and the level of the classroom are poorly conceptualised, rarely explicit, and even more rarely practically drawn.

The problems with the, historically, mostly ‘school level’ orientation of school improvements as judged against the literature are, of course, that:

- Within school variation by department within secondary school and by teacher within primary school is much greater than the variation between schools on their ‘mean’ levels of achievement and ‘value added’ effectiveness (Fitz-Gibbon, 1991);
- The effect of the teacher and of the classroom level in those multi-level analyses that have been undertaken, since the introduction of this technique in the mid-1980s, is probably three to four times greater than that of the school level (Muijs & Reynolds, 2011).

A classroom or 'learning level' orientation is likely to be more productive than a 'school level' orientation for achievement gains, for the following reasons:

- The classroom can be explored using the techniques of 'pupil voice' that are now so popular;
- The classroom level is closer to the student level than is the school level, opening up the possibility of generating greater change in outcomes through manipulation of 'proximal variables';
- Whilst not every school is an effective school, every school has within itself some classroom practice that is relatively more effective than its other practice. Many schools will have within themselves classroom practice that is absolutely effective across all schools. With a within school 'learning level' orientation, every school can benefit from its own internal conditions;
- Focussing on classroom may be a way of permitting greater levels of competence to emerge at the school level;
- There are powerful programmes (e.g. Slavin, 1996) that are classroom-based, and powerful approaches, such as peer tutoring and collaborative groupwork;
- There are extensive bodies of knowledge related to the factors that effective teachers use and much of the novel cognitive neuroscience material that is now so popular internationally has direct 'teaching' applications;
- There are techniques, such as lesson study, that can be used to transfer good practice, as outlined historically in *The Teaching Gap* (Stigler & Hiebert, 1999).

3.3.2 *Lesson Study and Collaborative Enquiry to Foster SI*

Much is made in this latter study of the professional development activities of Japanese teachers, who adopt a 'problem-solving' orientation to their teaching, with the dominant form of in-service training being the lesson study. In lesson study, groups of teachers meet regularly over long periods of time (ranging from several months to a year) to work on the design, implementation, testing, and improvement of one or several 'research lessons'. By all indications, report Stigler and Hiebert (1999),

lesson study is extremely popular and highly valued by Japanese teachers, especially at the elementary school level. It is the linchpin of the improvement process and the premise behind lesson study is simple: If you want to improve teaching, the most effective place to do so is in the context of a classroom lesson. If you start with lessons, the problem of how to apply research findings in the classroom disappears. The improvements are devised within the classroom in the first place. The challenge now becomes that of identifying the kinds of changes that will improve student learning in the classroom and, once the changes are identified, of sharing this knowledge with other teachers, who face similar problems, or share similar goals in the classroom. (p. 110)

It is the focus on improving instruction within the context of the curriculum, using a methodology of collaborative enquiry into student learning, that provides the usefulness for contemporary school improvement efforts. The broader argument is that

it is this form of professional development, rather than efforts at only *school* improvement, that provides the basis for the problem-solving approach to teaching adopted by Japanese teachers.

3.4 Building School Improvement Capacity

We noted earlier that conventional educational reforms may not have delivered enhanced educational outcomes because they did not affect school capacity to improve, merely assuming that educational professionals were able to surf the range of policy initiatives to good effect. Without the possession of ‘capacity,’ schools will be unable to sustain continuous improvement efforts that result in improved student achievement. It is therefore critical to be able to define ‘capacity’ in operational terms. The IQEA school improvement project, for example, demonstrated that without a strong focus on the internal conditions of the school, innovation work quickly becomes marginalised (Hopkins 2001). These ‘conditions’ have to be worked on at the same time as the curriculum on other priorities the school has set itself and are the internal features of the school, the ‘arrangements’ that enable it to get its work done (Ainscow et al., 2000). The ‘conditions’ within the school that have been associated with a capacity for sustained improvement are:

- A commitment to staff development
- Practical efforts to involve staff, students, and the community in school policies and decisions
- ‘Transformational’ leadership approaches
- Effective co-ordination strategies
- Serious attention to the benefits of enquiry and reflection
- A commitment to collaborative planning activity

The work of Newmann, King, and Young (2000) provided another perspective on conceptualising and building learning capacity. They argue that professional development is more likely to advance achievement for all students in a school, if it addresses not only the learning of individual teachers, but also other dimensions concerned with the organisational capacity of the school. They defined school capacity as the collective competency of the school as an entity to bring about effective change. They suggested that there are four core components of capacity:

- The knowledge, skills, and dispositions of individual staff members;
- A professional learning community – in which staff work collaboratively to set clear goals for student learning, assess how well students are doing, and develop action plans to increase student achievement, whilst being engaged in inquiry and problem-solving;
- Programme coherence – the extent to which the school’s programmes for student and staff learning are co-ordinated, focused on clear learning goals and sustained over a period of time;

- Technical resources – high quality curriculum, instructional material, assessment instrument, technology, workspace, etc.

Fullan (2000) notes that this four-part definition of school capacity includes ‘human capital’ (i.e. the skills of individuals), but he concludes that no amount of professional development of individuals will have an impact, if certain organisational features are not in place. He maintains that there are two key organisational features necessary. The first is ‘professional learning communities’, which is the ‘social capital’ aspect of capacity. In other words, the skills of individuals can only be realised, if the relationships within the schools are continually developing. The other component of organisational capacity is programme coherence. Since complex social systems have a tendency to produce overload and fragmentation in a non-linear, evolving fashion, schools are constantly being bombarded with overwhelming and unconnected innovations. In this sense, the most effective schools are not those that take on the most innovations, but those that selectively take on, integrate and co-ordinate innovations into their own focused programmes.

A key element of capacity building is the provision of in-classroom support, or in a Joyce and Showers term, ‘peer coaching’. It is the facilitation of peer coaching that enables teachers to extend their repertoire of teaching skills and to transfer them from different classroom settings to others. In particular, peer coaching is helpful when (Joyce, Calhoun, & Hopkins, 2009):

- Curriculum and instruction are the contents of staff development;
- The focus of the staff development represents a new practice for the teacher;
- Workshops are designed to develop understanding and skills;
- School-based groups support each other to attain ‘transfer of training’.

3.5 Studying the Interactions Between Schools, Homes, and Communities

Recent years have seen the SI field expand its interests into new areas of practice, although the acknowledgement of the importance of new areas has only to a limited degree been matched by a significant research enterprise to fully understand their possible importance.

Early research traditions established in the field encouraged the study of ‘the school’ rather than of ‘the home’ because of the oppositional nature of our education effectiveness community. Since critics of the field had argued that ‘schools make no difference’, we in EE, by contrast, argued that schools do make a difference and proceeded to study schools exclusively, not communities or families together with schools.

More recently, approaches, which combine school influences and neighbourhood/social factors in combination to maximise influence over educational achievement, have become more prevalent (Chapman et al., 2012). The emphasis is now

upon ‘beyond school’ rather than merely ‘between school’ influences. Specifically, there is now:

- A focus upon how schools cast their net wider than just ‘school factors’ in their search for improvement effects (Neeleman, 2019a), particularly, in recent years, involving a focus upon the importance of outside school factors;
- As EE research has further explored what effective schools do, the ‘levers’ these schools use have increasingly been shown to involve considerable attention to home and to community influences within the ‘effective’ schools;
- It seems that, as a totality, schools themselves are focussing more on these extra-school influences, given their clear importance to schools and given schools’ own difficulty in further improving the quality of already increasingly ‘maxed out’ internal school processes and structures; but this might also be largely context-dependent;
- Many of the case studies of successful school educational improvement, school change, and, indeed, many of the core procedures of the models of change employed by the new ‘marques’ of schools, such as the Academies’ Chains in the United Kingdom and Charter Schools in the United States, give an integral position to schools attempting to productively link their homes, their community, and the school;
- It has become clear that variance in outcomes explained by outside school factors is so much greater than the potential effects of even a limited, synergistic combination of school and home influences could be considerable in terms of effects upon school outcomes;
- The variation in the characteristics of the outside world of communities, homes, and caregivers itself is increasing considerably with the rising inequalities of education, income, and health status. It may be that these inequalities are also feeding into the maximisation of community influences upon schools and, therefore, potentially the mission of SI. At least, we should be aware of the growing gap between the haves and the have-nots (or, following David Goodhart, the somewheres and the anywheres) in many Western (European) countries and its possible influence on educational outcomes.

3.6 Delivering School Improvement Is Difficult!

Even accepting that we are clear on the precise ‘levers’ of school improvement, and we have already seen the complexity of these issues, it may be that the characteristics, attributes, and attitudes of those in schools, who are expected to implement improvement changes, may somewhat complicate matters. The work of Neeleman (2019a), based on a mixed-methods study among Dutch secondary school leaders, suggests a complicated picture:

- School improvement is general in nature rather than being specifically related to the characteristics of schools and classrooms outlined in research;
- School leaders' personal beliefs relate to connecting and collaborating with others, a search for moral purpose and the need to facilitate talent development and generate well-being and safe learning environments. Their core beliefs are about strong, value- driven, holistic, people-centred education, with an emphasis on relationships with students and colleagues. Rather than being motivated by the ambition to improve students' cognitive attainment, which is what school improvement and school improvers emphasize.
- School leaders interpret cognitive student achievement as a set of externally defined accountability standards. As long as these standards are met, they are rather motivated by holistic, development-oriented, student-centred, and non-cognitive ambitions. This is rather striking in light of current debates about the alleged influence of such standardized instruments on school practices, as critics have claimed that these instruments limit and steer practitioners' professional autonomy.
- Instead of concluding that school leaders are not driven by the desire to improve cognitive student achievement as commonly defined in EE research or enacted in standardized accountability frameworks, one could also claim that school leaders define or enact the notion differently. Rather than finding the continuous improvement of cognitive student achievement the holy grail of education, they seem more driven by the goal of offering their students education that prepares them for their future roles in a changing society. This interpretation implies more customized education with a focus on talent development and noncognitive outcomes, such as motivation and ownership. Such objectives, however, are seldom used as outcome measures in EE research or accountability frameworks.
- If evidence plays a role in school leaders' intervention decision-making, it is often used implicitly and conceptually, and it frequently originates from personalized sources. This suggests a rather minimal direct use of evidence in school improvement. The liberal conception of evidence that school leaders demonstrate is striking, all the more so, if one compares this interpretation to common conceptions of evidence in policy and academic discussions about evidence use in education. School leaders tend to assign a greater role to tacit knowledge and intuition in their decision-making than to formal or explicit forms of knowledge

In all, these findings raise questions in light of the ongoing debate about the gap between educational research and practice. If, on the one hand, school leaders are generally only slightly interested in using EE research, this would indicate the failure of past EE efforts. If, on the other hand, school leaders are indeed interested in using more EE evidence in their school improvement efforts, but insufficiently recognize common outcome measures or specific (meta-)evidence on their considered interventions, then we have a different problem. These questions require answers, if we want to bridge the gap between EE and SI and, thereby, strengthen school improvement capacity.

References

- Ainscow, M., Farrell, P., & Tweddle, D. (2000). Developing policies for inclusive education: a study of the role of local education authorities. *International Journal of Inclusive Education*, 4(3), 211–229.
- Boschloo, A., Krabbendam, L., Dekker, S., Lee, N., de Groot, R., & Jolles, J. (2013). Subjective sleepiness and sleep quality in adolescents are related to objective and subjective measures of school performance. *Frontiers in Psychology*, 4(38), 1–5. <https://doi.org/10.3389/fpsyg.2013.00038>
- Broekkamp, H., & Van Hout-Wolters, B. (2007). The gap between educational research and practice: A literature review, symposium, and questionnaire. *Educational Research and Evaluation*, 13(3), 203–220. <https://doi.org/10.1080/13803610701626127>
- Brown, C., & Greany, T. (2017). The evidence-informed school system in England: Where should school leaders be focusing their efforts? *Leadership and Policy in Schools*. <https://doi.org/10.1080/15700763.2016.1270330>
- Chapman, C., Muijs, D., Reynolds, D., Sammons, P., & Teddlie, C. (2016). *The Routledge international handbook of educational effectiveness and improvement: Research, policy, and practice*. London and New York, NY: Routledge.
- Chapman, C., Armstrong, P., Harris, A., Muijs, D., Reynolds, D., & Sammons, P. (Eds.). (2012). *School effectiveness and school improvement research, policy and practice: Challenging the orthodoxy*. New York, NY/London, UK: Routledge.
- Creemers, B., & Kyriakides, L. (2009). Situational effects of the school factors included in the dynamic model of educational effectiveness. *South African Journal of Education*, 29(3), 293–315.
- Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership*, 37(1), 15–27.
- Fitz-Gibbon, C. T. (1991). Multilevel modelling in an indicator system. In S. Raudenbush & J. D. Willms (Eds.), *Schools, pupils and classrooms: International studies of schooling from a multilevel perspective* (pp. 67–83). London, UK/New York, NY: Academic.
- Fullan, M. (2000). The return of large-scale reform. *Journal of Educational Change*, 1(1), 5–27.
- Galton, M. (1987). An ORACLE chronicle: A decade of classroom research. *Teaching and Teacher Education*, 3(4), 299–313.
- Hallinger, P., & Murphy, J. (1986). The social context of effective schools. *American Journal of Education*, 94(3), 328–355.
- Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.
- Hopkins, D. (2001). *Improving the quality of education for all*. London: David Fulton Publishers.
- Hopkins, D. (2007). *Every school a great school*. Maidenhead: Open University Press.
- Joyce, B. R., Calhoun, E. F., & Hopkins, D. (2009). *Models of learning: Tools for teaching* (3rd ed.). Maidenhead: Open University Press.
- Levin, B. (2004). Making research matter more. *Education Policy Analysis Archives*, 12(56). Retrieved from <http://epaa.asu.edu/epaa/v12n56/>
- Marzano, R. J. (2003). *What works in schools: Translating research into action*. Alexandria, VA: ASCD.
- May, H., Huff, J., & Goldring, E. (2012). A longitudinal study of principals' activities and student performance. *School Effectiveness and School Improvement*, 23(4), 415–439.
- Muijs, D., Harris, A., Chapman, C., Stoll, L., & Russ, J. (2004). Improving schools in socio-economically disadvantaged areas: A review of research evidence. *School Effectiveness and School Improvement*, 15(2), 149–175.
- Muijs, D., & Reynolds, D. (2011). *Effective teaching: Evidence and practice*. London, UK: Sage.
- Neeleman, A. (2019a). *School autonomy in practice. School intervention decision-making by Dutch secondary school leaders*. Maastricht, The Netherlands: Universitaire Pers Maastricht.

- Neeleman, A. (2019b). The scope of school autonomy in practice: An empirically based classification of school interventions. *Journal of Educational Change*, 20(1), 3155. <https://doi.org/10.1007/s10833-018-9332-5>
- Newmann, F., King, B., & Young, S. P. (2000). Professional development that addresses school capacity. Paper presented at American Educational Research Association Annual Conference, New Orleans, 28 April.
- Reynolds, D., Creemers, B. P. M., Stringfield, S., Teddlie, C., & Schaffer, E. (2002). *World class schools: International perspectives in school effectiveness*. London, UK: Routledge Falmer.
- Reynolds, D. (2010). *Failure free education? The past, present and future of school effectiveness and school improvement*. London, UK: Routledge.
- Reynolds, D., Sammons, P., de Fraine, B., van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, 25(2), 197–230.
- Robinson, V., Hohepa, M., & Lloyd, C. (2009). *School leadership and student outcomes: Identifying what works and why. Best evidence synthesis iteration (BES)*. Wellington, New Zealand: Ministry of Education.
- Scheerens, J. (2016). *Educational effectiveness and ineffectiveness. A critical review of the knowledge base*. Dordrecht, The Netherlands: Springer.
- Slavin, R. E. (1996). *Education for all*. Lisse, The Netherlands: Swets & Zeitlinger.
- Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press, Simon and Schuster, Inc.
- Stoll, L., & Myers, K. (1998). *No quick fixes*. London, UK: Falmer Press.
- Teddlie, C., & Stringfield, S. (1993). *Schools make a difference: Lessons learned from a 10-year study of school effects*. New York, NY: Teachers College Press.
- Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London, UK: Falmer Press.
- Vanderlinde, R., & van Braak, J. (2009). The gap between educational research and practice: Views of teachers, school leaders, intermediaries and researchers. *British Educational Research Journal*, 36(2), 299–316. <https://doi.org/10.1080/01411920902919257>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 4

The Relationship Between Teacher Professional Community and Participative Decision-Making in Schools in 22 European Countries



Catalina Lomos

4.1 Introduction

The literature on school effectiveness and school improvement highlights a positive relationship between professional community and participative decision-making in creating sustainable innovation and improvement (Hargreaves & Fink, 2009; Harris, 2009; Smylie, Lazarus, & Brownlee-Conyers, 1996; Wohlstetter, Smyer, & Mohrman, 1994). Many authors, beginning with Little (1990) and Rosenholtz (1989), indicated that teachers' participation in decision-making builds upon teacher collaboration and that the interaction of these elements leads to positive change and better school performance (Harris, 2009). Moreover, Carpenter (2014) indicated that school improvement points to a focus on professional community practices as well as supportive and participative leadership.

Broad participation in decision-making across the school is believed to promote cooperation and student development via valuable exchange regarding curriculum and instruction. Smylie et al. (1996) see a relevant and positive relationship, especially between participation in decision-making and teacher collaboration for learning and development, in the form of professional community. The authors consider that participation in decision-making may affect relationships between teachers and organisational learning opportunities due to increased responsibility, greater perceived accountability, and mutual obligation to respect the decisions made together.

Considering the desideratum of school improvement when identifying what factors facilitate better teacher and student outcomes (Creemers, 1994), the positive relationship between teacher collaboration within professional communities and teacher/staff participation in decision-making becomes of higher interest. The question that arises is whether this study-specific positive relationship identified can be

C. Lomos (✉)

Luxembourg Institute of Socio-Economic Research (LISER), Esch-sur-Alzette, Luxembourg
e-mail: Catalina.Lomos@liser.lu

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_4

41

considered universal and can be found across countries and educational systems. Therefore, the present study aims to investigate the following research questions across 22 European countries:

1. What is the relationship between professional community and participation in decision-making in different European countries?
2. Which of the actors involved in decision-making are most indicative of higher perceived professional community practices in different European countries?

In order to answer these research questions, the relationship between the two concepts needs to be estimated and compared across countries. Many authors, such as Billiet (2003) or Boeve-de Pauw and van Petegem (2012), have indicated how distorted cross-cultural comparisons can be when cross-cultural non-equivalence is ignored; thus, testing for measurement invariance of the latent concepts of interest should be a precursor to all country comparisons. The present chapter will answer these questions by applying a test for measurement invariance of the professional community latent concept as a cross-validation of the classical, comparative approach and will then discuss the impact of such a test on results.

4.2 Theoretical Section

4.2.1 Professional Community (PC)

Professional Community (PC) is represented by the teachers' level of interaction and collaboration within a school; it has been empirically established as relevant to teachers' and students' work (e.g. Hofman, Hofman, & Gray, 2015; Louis & Kruse, 1995). The concept has been under theoretical scrutiny for the last three decades, with the agreement that teachers are part of a professional community when they agree on a common school vision, engage in reflective dialogue and collaborative practices, and feel responsible for school improvement and student learning (Lomos, Hofman, & Bosker, 2012; Louis & Marks, 1998).

Regarding these specific dimensions of PC, Kruse, Louis, and Bryk (1995), "designated five interconnected variables that describe what they called genuine professional communities in such a broad manner that they can be applied to diverse settings" (Toole & Louis, 2002, p. 249). These five dimensions measuring the latent concept of professional community have been defined, based on Louis and Marks (1998) and other authors, as follows: *Reflective Dialogue* (RD) refers to the extent to which teachers discuss specific educational matters and share teaching activities with one another on a professional basis. *Deprivatisation of Practice* (DP) means that teachers monitor one another and their teaching activities for feedback purposes and are involved in observation of and feedback on their colleagues. *Collaborative Activity* (CA) is a temporal measure of the extent to which teachers engage in cooperative practices and design instructional programs and plans

together. *Shared sense of Purpose* (SP) refers to the degree to which teachers agree with the school's mission and take part actively in operational and improvement activities. *Collective focus or Responsibility for student learning* (CR) and a collective responsibility for school operations and improvement in general indicate a mutual commitment to student learning and a feeling of responsibility for all students in the school. This definition of PC has also been the measure most frequently used to investigate the PC's quantitative relationship with participative decision-making (e.g. Louis & Kruse, 1995; Louis & Marks, 1998; Louis, Dretzke, & Wahlstrom, 2010).

4.2.2 *Participative Decision-Making (PDM)*

The framework of participative decision-making as a theory of leadership practice has long been studied and has multiple applications in practice. Workers' involvement in the decisions of an organization has been investigated for its efficacy since 1924, as indicated by the comprehensive review of Lowin for the years between 1924 and 1968 (Conway, 1984). Regarding the involvement of educational personnel and the details of their participation, Conway (1984) characterizes their participation as "mandated versus voluntary", "formal versus informal", and "direct versus indirect". These dimensions differentiate the involvement of different actors, who could be involved in decision-making within schools. A few studies performed later, once school-based, decision-making measures had been implemented, such as Logan (1992) in the US state of Kentucky, listed principals, counsellors, academic and non-academic teachers and students as school personnel actively involved in decision-making.

When referring to participation in decision-making (PDM), specifically in educational organizations, Conway (1984) described the concept as an intersection of two major conceptual notions: *decision-making* and *participation*. Decision-making indicates a process, in which one or more actors determine a particular choice. Participation signifies "the extent to which subordinates, or other groups who are affected by the decisions, are consulted with, and involved in, the making of decisions" (Melcher, 1976, p. 12, in Conway, 1984).

Conway (1984) discusses the external perspective, which implies the participation of the broader community, and the internal perspective, which implies the participation of school-based actors. In many countries, including England (Earley & Weindling, 2010), the school *governors* are expected to have an important non-active leadership role in schools, more focused on "strategic direction, critical friendship and accountability" (p. 126), providing support and encouragement. The school *counsellor* has more of a supportive leadership role in facilitating the academic achievement of all students (Wingfield, Reese, & West-Olantunji, 2010) and enabling a stronger sense of school community (Janson, Stone, & Clark, 2009). *Teacher* participation can take the form of individual leadership roles for teachers or teacher advisory groups (Smylie et al., 1996). *Students* are also actors in

participative decision-making, especially when decisions involve the instructional process and learning materials. Students would need to discuss topics and learning activities with one another and their teachers to be informed for such decision-making; this increases the likelihood of collaborative interactions (Conway, 1984). Significantly, teachers have been identified as the most important actors, either formally or informally involved in participative decision-making; as such, reform proposals have recommended the expansion of teachers' participation in leadership and decision-making tasks (Louis et al., 2010).

4.2.3 The Relationship Between Professional Community and Participative Decision-Making

After following schools implementing participative decision-making with different actors involved, many studies found it imperative for teachers to interact if any meaningful and consistent sharing of information was to occur (e.g. Louis et al., 2010; Smylie et al., 1996). Moreover, they found that participative decision-making promotes collaboration and can bring teachers together in school-wide discussions. This phenomenon could limit separatism and increase interaction between different types of teachers (e.g. academic or vocational), especially in secondary schools (Logan, 1992). These studies also found that schools move towards mutual understanding through participation in decision-making, thus facilitating PC (p. 43). For Spillane, Halverson, and Diamond (2004), PC can facilitate broader interactions within schools. The authors have also concluded that “the opportunity for dialogue contributes to breaking down the school’s ‘egg-carton’ structure, creating new structures that support peer-communication and information-sharing, arrangements that in turn contribute to defining their leadership practice” (p. 27).

In conclusion, the relationship between professional community (PC) and actors of participative decision-making (PDM) has found to be significant and positive in different studies performed across varied educational systems (e.g. Carpenter, 2014; Lambert, 2003; Louis & Marks, 1998; Logan, 1992; Louis et al., 2010; Morrissey, 2000; Purkey & Smith, 1983; Smylie et al., 1996; Stoll & Louis, 2007). These findings support our expectation that this relationship is positive; PC and PDM mutually and positively influence each other over time, and this interaction creates paths to educational improvement (Hallinger & Heck, 1996; Pitner, 1988).

4.2.4 The Specific National Educational Contexts

Professional requirements to obtain a position as a teacher or a school leader vary widely across Europe. The 2013 report (Eurydice, 2013, Fig. F5, p. 118) describes the characteristics of participative decision-making, as well as other data, from

2011–2012 (relevant period for the present study) from pre-primary to upper secondary education in the studied countries.

From this report, we see that some countries share characteristics of participative decision-making; however, no typology of countries has yet been established or tested in this regard.

In most of the countries, participation is formal, mandated, and direct (Conway, 1984). More specifically, in countries, such as Belgium (Flanders) (BFL), Cyprus (CYP), the Czech Republic (CZE), Denmark (DNK), England (ENG), Spain (ESP), Ireland (IRL), Latvia (LVA), Luxembourg (LUX), Malta (MLT), and Slovenia (SVN), school leadership is traditionally shared among formal leadership teams and team members. Principals, teachers, community representatives and, in some countries, governing bodies all typically constitute formal leadership teams. For most, the formal tasks deal with administration, personnel management, maintenance, and infrastructure rather than with pedagogy, monitoring, and evaluation (Barrera-Osorio, Fasih, Patrinos, & Santibanez, 2009).

In other European countries, such as Austria (AUT), Bulgaria (BGR), Italy (ITA), Lithuania (LTU), and Poland (POL), PDM occurs as a combination of formal leadership teams and informal ad-hoc groups. Ad-hoc leadership groups are created to take over specific and short-term leadership tasks, complementing the formal leadership teams. For example, in Italy these leadership roles can be defined for an entire year, and in most countries, there is no external incentive to reward participation. Participation depends upon the input of teaching and non-teaching staff, such as parents, students, and the local community, through school boards or school governors, student councils and teachers' assemblies (p. 117). In these cases, participation is more active through collaboration and negotiation of decisions. In addition, the responsibilities of PDM range from administrative or financial to specifically pedagogical or managerial. In Malta, for example, the participative members focus more on administrative and financial matters, while in Slovenia, the teaching staff creates a professional body that makes autonomous decisions about program improvement and discipline-related matters (p. 117).

In Nordic countries, such as Estonia (EST), Finland (FIN), Norway (NOR), and Sweden (SWE), schools make decisions about leadership distribution with the school leader having a key role in distributing the participative responsibilities. The participating actors are mainly the leaders of the teaching teams that implement the decisions.

One unique country, in terms of PDM, is Switzerland (CHE), where no formal distribution of school leadership and decision-making takes place.

In terms of the presence of professional community, Lomos (2017) has comparatively analyzed the presence of PC practices in all the European countries mentioned above. It was found that teachers in Bulgaria and Poland perceive significantly higher PC practices than the teachers in all other participating European countries. After Bulgaria and Poland, the group of countries with the next-highest, albeit significantly lower factor mean includes Latvia, Ireland, and Lithuania; teachers' PC perceptions in these countries do not differ significantly. The third group of countries with significantly lower PC latent scores is comprised of Slovenia, England,

and Switzerland, followed in the middle by a larger group of countries, which includes Italy, Spain, Sweden, Norway, Finland, Estonia, and Slovakia, and followed lower by Malta, Cyprus, the Czech Republic, and Austria. Belgium (Flanders) (BFL) proves to have the lowest mean of the PC factor; it is lower than those of 19 other European countries, excluding Luxembourg and Denmark, which have PC means that do not differ significantly from that of Belgium (Flanders) (BFL).

Considering the present opportunity to study these relationships across many countries, it is important to know which decision-making actors most strongly indicate a high level of PC and whether different patterns of relationships appear for specific actors in different countries. While the TALIS 2013 report (OECD, 2016) treated the shared participative leadership concept as latent and investigated its relationship with each of the five PC dimensions separately, the present study aims to go a step further by clarifying what actors involved in decision-making prove most indicative of higher PC practices in general. Treating PC as one latent concept allows us to formulate conclusions about the effect of each actor involved in PDM on the general collaboration level within schools rather than on each separate PC dimension. To formulate such conclusions at the higher-order level of the PC latent concept, a test of measurement invariance is necessary, which will be presented later in this chapter.

Considering the exploratory nature of this study, in which the relationship between the PC concept and PDM actors will be investigated comparatively across many European countries, no specific hypotheses will be formulated. The only empirical expectation that we have across all countries, based on existing empirical evidence, is that this relationship is positive; PC and PDM actors mutually and positively influence each other.

4.3 Method

4.3.1 Data and Variables

The present study uses the European Module of the International Civic and Citizenship Education Study (ICCS 2009) performed in 23 countries.¹ The ICCS 2009 evaluates the level of students' civic knowledge in eighth grade (13.5 years of age and older), while also collecting data from teachers, head teachers, and national

¹The countries in the European module and included in this study are: Austria (AUT) N teachers = 949, Belgium (Flemish) (BFL) N = 1582, Bulgaria (BGR) N = 1813, Cyprus (CYP) N = 875, the Czech Republic (CZE) N = 1557, Denmark (DNK) = 882, England (ENG) N = 1408, Estonia (EST) N = 1745, Finland (FIN) N = 2247, Ireland (IRL) N = 1810, Italy (ITA) N = 2846, Latvia (LVA) N = 1994, Liechtenstein (LIE) N = 112, Lithuania (LTU) N = 2669, Luxembourg (LUX) N = 272, Malta (MLT) N = 862, Norway (NOR) N = 482, Poland (POL) N = 2044, Slovakia (SVK) N = 1948, Slovenia (SVN) N = 2698, Spain (ESP) N = 1934, Sweden (SWE) N = 1864, and Switzerland (CHE) N = 1416. Greece and the Netherlands have no teacher data available.

representatives. When answering the specific questions, teachers - the unit of analysis in this study - also indicated their perception of collaboration within their school, their contribution to the decision-making process, and students' influence on different decisions made within their school. In each country, 150 schools were selected for the study; from each school, one intact eighth-grade class was randomly selected and all its students surveyed. In small countries, with fewer than 150 schools, all qualifying schools were surveyed. Fifteen teachers teaching eighth grade within each school were randomly selected from all countries; in schools with fewer than 15 eighth-grade teachers, all eighth-grade teachers were selected (Schulz, Ainley, Fraillon, Kerr, & Losito, 2010). Therefore, the ICCS unweighted data from 23 countries include more than 35,000 eighth-grade teachers, with most countries having around 1500 participating teachers (see Footnote 1 for each country's unweighted teacher sample size). The unweighted sample size varied from 112 teachers in Liechtenstein to 2846 in Italy based on the number of schools in each country and the number of selected teachers ultimately answering the survey.

In the ICCS 2009 teacher questionnaire, five items were identified as an appropriate measurement of the Professional Community latent concept in this study. Namely, the teachers were asked how many teachers in their school during the current academic year:

- Support good discipline throughout the school even with students not belonging to their own class or classes? (Collective Responsibility/CR)
- Work collaboratively with one another in devising teaching activities? (Reflective Dialogue/RD)
- Take on tasks and responsibilities in addition to teaching (tutoring, school projects, etc.)? (Deprivatisation of Practice/DP)
- Actively take part in <school development/improvement activities>²? (Shared sense of Purpose/SP)
- Cooperate in defining and drafting the <school development plan>? (Collaborative Activity/CA)

These items, presented in the order in which they appeared in the original questionnaire, refer to teacher practices embedded into the five dimensions of PC. The five items were measured using a four-point Likert scale that went from "all or nearly all" to "none or hardly any". For the analysis, all indicators were inverted in order to interpret the high numerical values of the Likert scale as indicators of high PC participation. Around 2.5% of data were missing across all five items on average across all countries. Most countries had a low level of missing data – only 1–2% – and the largest amount of missing data was 5%. No school or country completely lacked data. Any missing data for the five observed variables of the latent professional community concept were considered to be missing completely at random, and deletion was performed list-wise.

²The signs <...> mark country-specific actions, subject to country adaptation.

Participative decision-making was also measured through five items indicating the extent to which different school actors contribute to the decision-making process. First, three items measure how much the teachers perceive that the following groups contribute to decision-making:

- Teachers
- School Governors
- School Counsellors

Two additional items measure how much teachers perceive students' opinions to be considered when decisions are made about the following issues:

- Teaching and learning materials
- School rules

These five items were measured on a four-point Likert scale, which ranged from "to a large extent" to "not at all". For the analysis, all indicators were inverted in order to interpret the high-numerical values of the Likert scale as an indication of high involvement. The amount of missing data varied across the five items; about 1% of data regarding teacher involvement and consideration of students' opinions was missing across all the countries. On the question of school governors' involvement, about 11% of the data were missing across all countries (the question was not applicable in Austria and Luxembourg; 10% of missing cases were found for this question in Sweden and Switzerland). Moreover, 15% of missing cases were found on average for the item on school counsellors' involvement (the question was not applicable in Austria, Luxembourg, and Switzerland; 10% of missing cases were found for this question in Bulgaria, Estonia, Lithuania, and Sweden). The missing data were deleted list-wise, but the countries with more than 10% missing cases were flagged for caution in the results' graphical representations, when interpreting these countries' outcomes due to the possible self-selection by the teachers, who actually answered the questions.

4.3.2 Analysis Method

First, the scale reliability and the factor composition of the PC scale were tested across countries and in each individual country through both reliability analysis (Cronbach α for the entire scale) and factor analysis (EFA with Varimax rotation). Conditioned on the results obtained, the PC scale was built as the composite scale score, and the relationship of the scale with each item measuring PDM was investigated through correlation analysis. The level of significance was considered one-tailed since positive relationships were expected. The five items measuring PDM were correlated individually with the PC scale in an attempt to disentangle what PDM aspect within schools matters most to such collaborative practices across all countries. Considering the multitude of tests applied, the Holm-Bonferroni correction indicates in this case the level of $p < .002$ ($\alpha/21$) as the p -value to reject the

null-hypothesis; the correlation bars respecting this condition are indicated with a bold pattern in the results section (see Figures).

To account for the specifics of the ICCS 2009 data, the IEA IDB Analyzer program (IEA, 2017) was used to perform all analyses, accounting for the specifics of the data through stratification, weights, and clustering adjustments, allowing us to make valid conclusions at the teacher level. These adjustments correct for the sampling strategy across countries and for the nested character of the data. Same data-specific adjustments were applied to any analyses performed in SPSS (SPSS statistics 24), such as the reliability analysis and factor analysis.

Considering that we are comparing correlation coefficients with the latent PC concept across countries, it is important to consider the equivalence of the measurement model for latent concepts in all groups. This will ensure that the associations found are in fact determined by the relationship between the concepts of interest and not by non-equivalent measurement models (Meuleman & Billiet, 2011). Therefore, a sensitivity check was performed in this chapter. First, as a cross-validation of the results obtained, the established and presented correlation coefficients were compared with the ones obtained applying the Multiple-Group Confirmatory Factor Analysis (MGCFA) method and taking into consideration the level of measurement metric invariance of the latent PC concept across all countries. The traditional MGCFA applied here for this cross-validation indicates that relationships with latent concepts can be validly compared across groups, if the latent concept has the same factor structure in all groups (configural invariance) and if the factor loadings of the measurement model are equal in all groups (metric invariance) (e.g. Meuleman & Billiet, 2012). For this chapter, the level of model fit in terms of metric invariance for the latent PC concept will be presented; the difference in the correlations obtained with the two methods (measurement model, either considered or not considered) will be discussed in terms of their implications on the presented results and interpretation. The Mplus program (Mplus 7.31) was used to perform the sensitivity analysis presented later in this chapter with all specific data adjustments applied (weights, strata, and clustering).

Further sensitivity checks of the relationships presented in this chapter were performed to test the robustness of the results. More specifically, the correlation coefficients obtained were corrected for different teachers' demographic characteristics (age, gender, teaching experience, subject taught in the current school, and other school responsibilities besides teaching) to make sure that the relationships presented are not spurious due to such variables. Finally, checks for linear relationships were performed as well, considering that all variables in this study were measured using four-point Likert scales.

4.4 Results

The results section will follow the order of the research questions, first presenting the relationships and their direction from each country while considering, which decision-making actor is indicative of high PC presence. Considering the exploratory character of this analysis, the correlation coefficients in all countries will be comparatively presented, and the most relevant results will be discussed.

The reliability analysis of the PC scale indicated satisfactory results across all countries ($\alpha = .78$, $N = 35,897$) and also in each individual country, with Cronbach α values ranging from .72 in Estonia to .87 in Luxembourg. Factor analysis indicated a one-factor structure across all countries with factor loadings higher than .68 showing also a one-factor structure in each country, excluding Estonia, where a two-factor solution, achieved by separating the first three and the last two PC items, fits better. However, the PC concept shows a satisfactory reliability level ($\alpha = .72$) in Estonia, indicating that we can keep this country in the analysis using the one-factor approach. Liechtenstein, did not show a satisfactory reliability and factor analysis result, so it was excluded from further analyses, leaving 22 European countries. For all other countries, the evidence presented here constitutes the basis of our confidence in creating the composite score for the PC concept and to use it for the following correlation analyses.

4.4.1 Professional Community and Participative Decision-Making

The following three Figures present the relationships measured between PC and the perceived involvement in decision-making of the teachers, school governor, and school counsellor.

In Fig. 4.1, we see a significant and positive correlation between PC and teacher decision-making in all countries with values ranging from $r = .23$ in Denmark

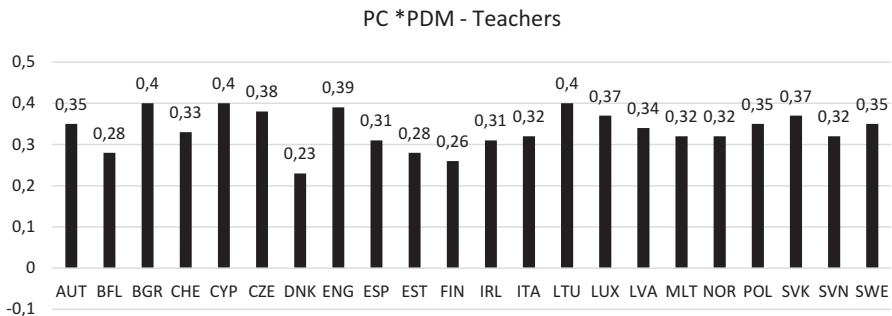


Fig. 4.1 PC and PDM – Teachers’ contribution to decision-making

Notes: Correlation coefficients obtained using the ICCS 2009 Teacher data, $N = 35,490$. The vertical X-line indicates the correlation coefficient for each country on a scale from -1.00 to $+1.00$; the horizontal Y-line indicates the country correlation bars in alphabetical order. All relationships are significant at the one-tailed value $p < .001$

(DNK) and $r = .26$ in Finland (FIN) to $r \approx .38$ in the Czech Republic (CZE) and England (ENG) and $r = .40$ in Bulgaria (BGR), Cyprus (CYP), and Lithuania (LTU). This outcome confirms previous empirical evidence that, when teachers are highly involved in their school’s decision-making process, they also perceive higher levels of participation in PC in their school; most countries have an r -value higher than $.30$.

In England, teachers are remunerated for some distribution of leadership functions, and for that, teachers need to manage pupils’ development along the curriculum (Eurydice, 2013). In Bulgaria, teachers receive additional points if they are involved in leading particular teams, and this can increase their payment, while in Cyprus, many teachers hold a Master’s degree in Leadership and Administration (Eurydice, 2013). However, in Finland, the school leader may or may not establish teams of teachers with leadership roles, and these teams may be disbanded in a flexible way based on the school’s interests (Eurydice, 2013).

The results are a bit different in Fig. 4.2, where we see that the relationship between PC and school governor decision-making is positive and statistically significant in all countries but with lower effect sizes, from $r \approx .10$ in Bulgaria (BGR), Spain (ESP), and Slovakia (SVK) to $r = .35$ in Poland (POL) and $r = .41$ in Lithuania (LTU). In all countries, a perception of high PC participation is not strongly related with a perception of school governors’ involvement in decision-making. This finding seems to indicate that having non-teaching staff involved in decision-making and assuming a more formal leadership role does not associate strongly with a high collaborative climate, as perceived by the teachers; the strength of the relationship varies considerably between countries.

In terms of general PDM within schools at the system level, much of the choice regarding who should be involved in decisions, and to what extent, remains with the school leaders in the countries studied. In Poland, the actors leading informal leadership teams are rewarded with merit-based allowances; this is also true of Lithuania, where there are no top-level incentives for distributing decision-making, so the initiative rests with the school leader (Eurydice, 2013).

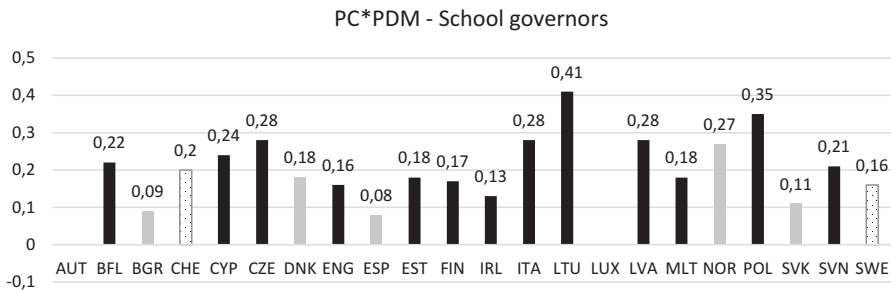


Fig. 4.2 PC and PDM – School governors’ contribution to decision-making

Notes: Correlation coefficients obtained using the ICCS 2009 Teacher data, $N = 31,439$. The vertical X-line indicates the correlation coefficient for each country on a scale from -1.00 to $+1.00$; the horizontal Y-line indicates the country correlation bars in alphabetical order. Relationships are significant at the one-tailed value $p < .001$; the lighter bars indicate a $p < .05$ level; the pattern-filled bars indicate more than 10% missing answers to this PDM question; missing bars indicate that the question was not asked in these countries

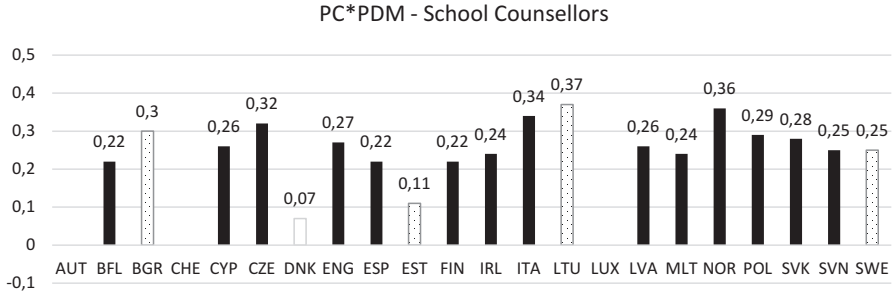


Fig. 4.3 PC and PDM – School Counsellors’ contribution to decision-making
Notes: Correlation coefficients obtained using the ICCS 2009 Teacher data, $N = 30,224$. The vertical X-line indicates the correlation coefficient for each country on a scale from -1.00 to $+1.00$; the horizontal Y-line indicates the country correlation bars in alphabetical order. All relationships are significant at the one-tailed value $p < .001$; the empty bars indicate a non-significant relationship; the pattern-filled bars indicate more than 10% missing cases for this PDM question; missing bars indicate that the question was not asked in these countries

The same varying relationship across countries can be noted in Fig. 4.3, where the school staff perceived as involved in decision-making is the school counsellor – in most countries, this is the student or educational/vocational career counsellor, psychologist, or social teacher.

One can see that in most countries, higher perceived PC is associated with higher perceived participation of school counsellors in decision-making; the majority shows a coefficient higher than $r = .22$, only Estonia (EST) is lower, and the data are not significant in Denmark (DNK). It is noteworthy that Lithuania (LTU), Poland (POL), Norway (NOR), the Czech Republic (CZE), Latvia (LVA), and Italy (ITA) are the countries with the strongest relationships between PC practices and the involvement of the school counsellor and, previously, the school governor in decision-making; these two relationships differ only for Bulgaria (BGR) and Slovakia (SVK) (see Figs. 4.2 and 4.3).

We also expected a positive relationship between the consideration of students’ opinions in decision-making and teachers’ PC participation, particularly when teachers cooperate to define the vision of the school and collaboratively take part in deciding what is best for their students. A positive and significant relationship between PC practices and the consideration of students’ opinions in decisions made about teaching and learning materials can be seen in Fig. 4.4.

In Fig. 4.4, the majority of coefficients is higher than $r = .20$, with lower ones only in Austria (AUT), Switzerland (CHE), Spain (ESP), Denmark (DNK), and Malta (MLT). In Austria, there are many pilot projects supporting the redistribution of tasks among formal and informal leadership teams, especially geared towards teachers but not necessarily students; meanwhile, Switzerland was reported as having no formally shared decision-making (Eurydice, 2013).

In terms of student opinions being considered when defining school rules, Fig. 4.5 depicts its relationship with PC as positive and relatively strong in all countries; again, most correlation coefficients are higher than $r = .20$. Some of the same

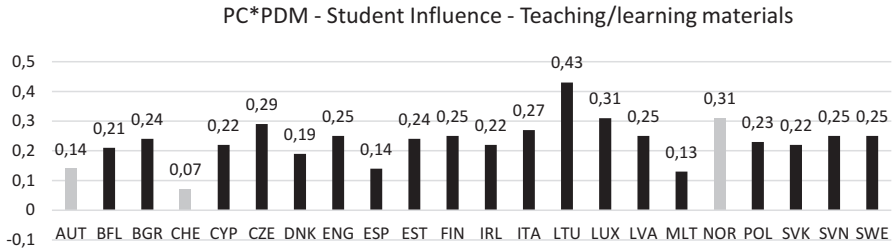


Fig. 4.4 PC and PDM – Student Opinions considered for Teaching and Learning materials
Notes: Correlation coefficients obtained using the ICCS 2009 Teacher data, $N = 35,105$. The vertical X-line indicates the correlation coefficient for each country on a scale from -1.00 to $+1.00$; the horizontal Y-line indicates the country correlation bars in alphabetical order. All relationships are significant at the one-tailed value $p < .001$; the lighter bars indicate a $p < .05$ level

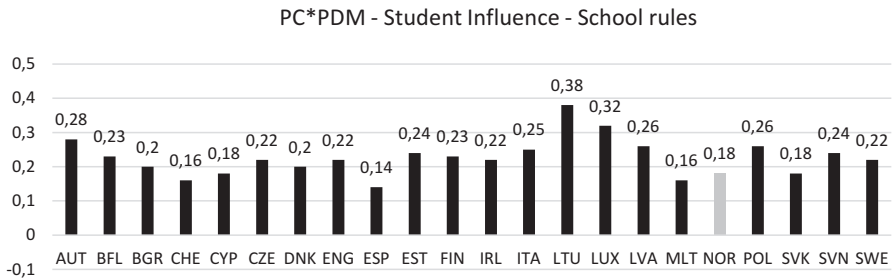


Fig. 4.5 PC and PDM – Student Opinions considered for School Rules
Notes: Correlation coefficients obtained using the ICCS 2009 Teacher data, $N = 35,105$. The vertical X-line indicates the correlation coefficient for each country on a scale from -1.00 to $+1.00$; the horizontal Y-line indicates the country correlation bars in alphabetical order. All relationships are significant at the one-tailed value $p < .001$; the lighter bars indicate a $p < .05$ level

countries have a lower r coefficient, such as Cyprus (CYP), Norway (NOR), and Slovakia (SVK), followed by even a lower r coefficient of Switzerland (CHE), Spain (ESP) and Malta (MLT). In general, in all countries, teachers agree that if they perceive their school as having a high level of participation in collaboration among teachers, they also perceive a high consideration of student opinions in defining school rules, and vice versa.

4.4.2 Sensitivity Checks

The results presented here have been cross-validated through three sensitivity checks, all of which concern the decisions made at the beginning of the study.

The first sensitivity check addresses the importance of the measurement metric invariance level of the latent PC concept and the comparison of its relationships

with PDM across the 22 groups. The traditional Multiple-Group Confirmatory Factor Analysis (MGCFA) indicates that relationships with latent concepts can be validly compared across groups, if the latent concept has the same factor structure in all groups (configural invariance) and if the factor loadings of the measurement model are equal in all groups (metric invariance) (e.g. Meuleman & Billiet, 2012). In Mplus, the metric invariance model³ within MGCFA was run; it showed a satisfactory model fit after freely estimating the factor loading for Switzerland’s Reflective Dialogue item, as recommended by the Model Modification Indices in JRule for Mplus (Sarlis, Satorra, & Van der Veld, 2009; Van der Veld & Sarlis, 2011), (CFI = .956, RMSEA = .066, ΔCFI = 1.0011, ΔRMSEA = 1.0011 compared to Full Metric Invariance, N = 35,897). Taking the test for metric invariance and its adjustments into consideration, the PC latent concept was correlated with each item of the PDM concept. In all countries, the correlation coefficients obtained by considering the metric measurement invariance testing were relatively higher than those obtained without considering the measurement model. The differences between the correlation coefficients for the two approaches ranged between .01 and up to .09 points (not tested for significant differences). These small differences found in the two approaches of estimating the relationship of PC and PDM involving teachers are presented in Fig. 4.6. Considering that the significance level of the relationships did not change in the present study and taking into account the relatively large sample size in each country, we have opted for the simpler approach, which does not consider the measurement invariance model of the latent PC concept when presenting

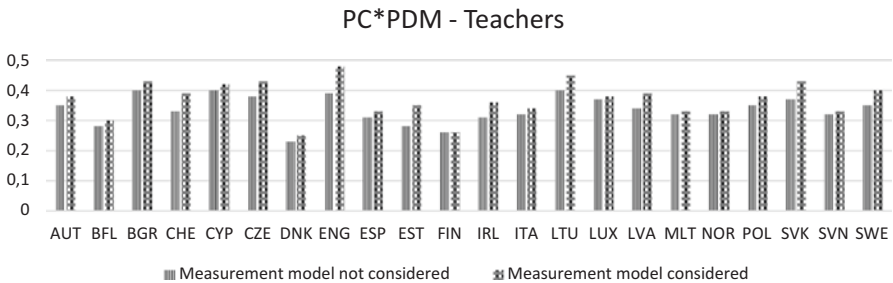


Fig. 4.6 PC and PDM – Teachers’ contribution to decision-making within schools – comparing correlation coefficients using two approaches in terms of measurement model considered
Notes: Correlation coefficients obtained using the ICCS 2009 Teacher data, N = 35,490. The vertical X-line indicates the correlation coefficient for each country on a scale from –1.00 to +1.00; the horizontal Y-line indicates the country correlation bars in alphabetical order. No label values were indicated to facilitate the easy reading of the figure, but the author can provide them

³The Full Metric Invariance Model within MGCFA was run, including a total of 7 corrections in terms of allowed error term correlations between 2 items (2 such error term correlations in Austria, Ireland, and England, and 1 in Estonia) as required by the individual Confirmatory Factor Analysis (CFA) models, ran in each individual country, and by an a-priori satisfactory model fit for the full configural measurement invariance model. The model fit for the Full Configural Invariance Model was satisfactory (CFI = .966, RMSEA = .079, N = 35,897).

the previous results. However, other studies should at least cross-validate their results by considering the measurement invariance model of latent concepts when comparing correlation coefficients; this will establish whether their relationships of interest are meaningful and supported by a satisfactory measurement metric invariance model across all groups.

The second sensitivity check addresses the relationships of interest and the risk of being spurious on demographic variables. It is possible that both the teachers' perception of PC and PDM practices are influenced by their gender, age, main subject taught (mathematics, languages, science, human sciences, or other subjects), or other roles within the school (member of the school council, assistant principal, department leader, guidance counsellor, or district representative) (e.g. Hulpia, Devos, & Rosseel, 2009a; Wahlstrom & Louise, 2008). To cross-validate the results, we have considered these variables alone and in different combinations in the correlation analyses performed. In all cases, the relationships stayed significant, and the size of the correlation coefficients did not change dramatically, i.e. increasing or decreasing by .05 points at most. Being a female, teaching mathematics, and being part of the school council triggered the correlation coefficient to change from .02 to .05 in some countries, such as Luxembourg (the country with the smallest sample size), but there was no change in the significance of the relationship.

The third sensitivity check addresses the decision of treating the observed items and the PC scale as continuous with all items being measured by four-point Likert scales. To cross-validate this decision, we investigated the distribution of the cases across the categories of all variables and in all countries. Across all countries, all observed variables had a lower number of responses for the lowest category ("none or hardly any" and "not at all"), with the exception of the PDM feature of students' influence on teaching and learning materials, which had a low response number for its highest category ("to a large extent"). In each case, we have merged each low- or high-response category with its closest neighboring category, creating variables with three categories each. The cross-tabulations, which were run across all countries and in each individual country, supported the expectation of a linear relationship.

4.5 Conclusion and Discussion

Returning to the research questions, Professional Community (PC) practices proved to be significantly and positively related to Participative Decision-Making (PDM) practices in all 22 European countries. Moreover, some actors, involved in PDM practices within schools, were more indicative of PC practices in all 22 countries, while other actors were relevant only in some countries.

All PDM features were positively and significantly related to PC practices in all countries; this is in accordance with the previous empirical evidence indicating that in schools, where such PDM structures are present, with teachers and other actors

involved in decision-making, there is also a higher presence of PC practices (Carpenter, 2014).

However, some school actors' involvement in decision-making is more indicative of the presence of PC practices than that of other actors. More specifically, the data prove that teachers' perception of high PC correlates the strongest with high levels of teacher involvement in decision-making. Furthermore, across all countries, more than 50% of the teachers, who perceived high levels of teacher involvement in decision-making, also perceived a strong presence of teacher professional community practices. This relationship proved weaker in Denmark, however, where weak PC practices were reported by those teachers, who perceived low teacher involvement in decision-making, and also by those, who perceived moderate teacher involvement. Moreover, even those teachers, who perceived high teacher involvement in decision-making in Denmark, mostly reported only a moderate presence of teacher PC practices. This might be influenced by a low teacher-perceived presence of professional community practices on average across schools in Denmark in the 2009 ICCS data, which also applies in Flanders (Belgium) (Lomos, 2017) and Estonia.

The degree of other actors' involvement in decision-making also has a positive relationship with the presence of PC practices, but the intensity of this relationship varies more widely across countries, sometimes being consistent with specific, formal PDM practices in different national educational contexts, as presented in the theoretical section.

In terms of school governors' involvement in decision-making, the size of the correlation coefficient in Bulgaria, Spain, and Slovakia was surprisingly low. Upon closer investigation of the distribution of responses, it became apparent that in these three countries, 90% of the teachers perceive the school governor to be largely involved in decision-making; the size of the correlation coefficient is, therefore, impacted by the lack of discrimination within this variable. This distribution of answers could be expected, considering that in these countries, the PDM is formal and traditionally shared among structured leadership teams and team members. In terms of the school counselors' involvement in decision-making, it can be noted that the majority of these relationships have a correlation coefficient larger than .20; it is lower only in Estonia, and it is not statistically significant in Denmark. In Denmark, 76% of the teachers, who answered this question, indicated that the school counselor is not involved in decision-making; the analysis shows no clear relationship in this country. In Estonia, only 7% of the teachers, who answered this question, indicated that the school counselor is highly involved in decision-making; most responses indicate no involvement. To conclude, the high involvement in decision-making of the school governor and school counselor in each country relates positively with a high perceived participation in professional community activities; however, this conclusion is perturbed in some countries by the formal and national regulations precisely defining the role and the attributions of such formal leader-followers within schools.

In terms of students' involvement in student-related decision-making and the presence of professional community practices, there is not much empirical

evidence, on which to base our expectations. From the TALIS 2013 cross-countries study (OECD, 2016), it is known that principals perceive low student participation in decision-making in countries, such as Italy, the Slovak Republic, Spain, and the Czech Republic, and a high student participation in Latvia, Poland, Estonia, Norway, England, and Bulgaria, but not much evidence is available on its relationship with teacher professional community practices. In our study, we found that the consideration of students' opinions regarding school rules is positively related to participation in teachers' PC practices; this relationship varies in strength across countries. A similar pattern of relationship can be seen between the consideration of students' opinions of teaching and learning materials, as summarized here. In both cases, student participation – in decisions about school rules and about teaching and learning materials, in Lithuania and Luxembourg have the strongest relation with PC presence, while in Spain, Malta, and Switzerland have the weakest one. The case of Luxembourg is interesting, since it has on average a predominant low perception of professional community practices in schools (Lomos, 2017) and a low perception of student influence on teaching and learning materials and school rules, based on teachers' answers in the ICCS 2009 data. This indicates that most teachers perceive their school as having either both collaborative practices and student influence on decision-making or neither of the two. High degrees of student influence on teaching and learning materials seems to be especially characteristic of schools with a supportive, collaborative, and common-vision environment. In the cases of Spain and Switzerland, the weak relationship could be determined by the fact that most teachers perceived on average a lack of students' influence on teaching and learning materials and school rules, independently of their perceived level of PC practices. The cases of Austria and Norway are unique, showing a stronger correlation of PC practices with one of the PDM features of student influence and a weaker correlation with the other. This may be influenced by the fact that one of the PDM features is present to a much larger extent than the other or is more strongly supported by the respective national educational policies.

Regarding the issue of measurement invariance when comparing relationships of latent concepts across countries, the aim is to test whether such latent concepts can be measured by the observed indicators at hand in each country (configural invariance) and, especially, to test whether they are measuring the same construct the same way across different countries (metric invariance). In this study, we found that the correlation coefficients have relatively larger values, when the metric measurement model is considered - however, with no change in the significance of the results in the different countries. For future studies, comparing relationships of latent concepts across groups implies performing and adjusting for a satisfactory measurement model fit. It is suggested that future research at least cross-validates the results obtained without invariance testing, as is the approach here.

4.5.1 *Limitations and Future Research*

One methodological limitation is related to the design of the ICCS data; the aim of this large-scale study is to explain students' civic knowledge, attitudes, and behaviours toward the end of compulsory education. This implied that only eighth-grade teachers were randomly selected to participate in each school, reflecting, however, upon the practices of all their colleagues in their school.

A second, related limitation relates to the method by which the concepts of interest were measured by the ICCS teacher questionnaire. We were only able to capture who participated in decision-making and to what extent, but not exactly what the tasks and roles of these actors were. Hulpia et al. (2009a) identified different roles and tasks of followers when assuming leadership roles, which have an important impact on the measured outcomes. Moreover, Harris (2009) pointed out that when too many leaders are present, this could negatively affect team outcomes due to inconsistencies in responsibilities and roles or conflicting priorities and objectives. However, we are not able to account for these factors here. We only focused on the actors involved in decision-making and neither on the type of relationship nor on the quality of outcomes determined by this relationship. Kennedy, Deuel, Nelson, and Slavitt (2011) also identified several important attributes of participative leadership that would support the development of strong school communities and teacher collaboration, which we were not able to assess in order to understand what could determine the positive association found.

Following the same line of reasoning, the five dimensions of the PC concept have been measured with only one item each, while some previous studies used three or more items per dimension. Moreover, some of the items are proxies of the dimensions of interest, such as the item measuring deprivatisation of practice. This dimension is measured by teachers' willingness to take on additional tasks besides teaching, such as tutoring or school projects, which could require some deprivatisation of individual practice.

Another limitation of the present study is determined by the decision of considering the PC and PDM practices as teacher practices, expressed through teacher perceptions of school practices. The unit of analysis here is the teacher, and the same-school dependency of their answers has been corrected when obtaining the results. The interest of the present study is to grasp the relationship at the teacher level, but future research could consider these characteristics as school-based and investigate their impact at the school level as well, using a multilevel data analysis approach. The work of Scherer and Gustafsson (2015) could be applicable, especially when building more complex multilevel structural equation models with cross-level interactions; new research could consider PC and PDM as attributes of teachers or/and of schools, depending on the conceptualization and the theoretical relationships of interest. When considering the concepts as school characteristics, it would be relevant to account for the possible effects of other school characteristics, such as size, organization, complexity of environment, structural arrangement, and level of school performance (Hulpia, Devos, & Rosseel, 2009b; Scott, 1995 in

Spillane et al., 2004) and possibly social composition or community context. Louis, Mayrowetz, Smiley, and Murphy (2009) have also pointed out that the size of the school and the number of departments within a secondary school can affect the creation and quality of the relationship investigated. Such a comprehensive approach would require multilevel data analysis, which would also provide the within- and between-levels of variance.

Future studies could also investigate whether the measured relationships change over time at the macro-level by using the cross-sectional ICCS data measured in 1999, 2009, and 2016 for the countries available. However, to grasp how these relationships change over time at micro-level, longitudinal teacher data would be necessary. Such longitudinal teacher data would also allow researchers to dive into the causal relationships and understand how these concepts influence each other over time, thus creating paths to improve learning (Hallinger & Heck, 1996; Pitner, 1988).

Future research could focus on many aspects of the cross-country relationships identified. One interesting approach could be to explain why these relationships differ in intensity across countries. Future studies could try to classify the countries by European region; by the distinction made by Hofstede's classification (2001) between 'collectivist' and 'individualist' cultures (with Ning, Lee, and Lee (2015) arguing that knowledge-sharing and collaboration could be higher in collectivist countries); by level of students' success expressed comparatively across countries in large-scale assessment studies' results (e.g. the Programme for International Student Assessment (PISA); or others); by the type of educational system according to the degree of participative and collaborative practices among educational actors or the amount of investment in professional collaborative practices (Eurydice, 2013; Muijs, West, & Ainscow, 2010); by the within-country variation (data permitting), keeping in mind that larger European countries, such as Italy or Spain, might have different PDM policies between regions; and by other criteria concerning countries and educational systems. Understanding why countries align or differ in the relationships between school capacities and processes would help advance school effectiveness literature and its empirical explanations.

References

- Barrera-Osorio, F., Fasih, T., Patrinos, H. A., & Santibanez, L. (2009). Decentralized decision-making in schools. In *The theory and evidence on school-based management*. Washington, DC: The World Bank. Retrieved from World Bank http://siteresources.worldbank.org/EDUCATION/Resources/278200-1099079877269/547664-1099079934475/547667-1145313948551/Decentralized_decision_making_schools.pdf
- Billiet, J. (2003). Cross-cultural equivalence with structural equation modelling. In J. Harkness, F. Van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 247–265). New York, NY: Wiley.
- Boeve-de Pauw, J., & van Petegem, P. (2012). Cultural differences in the environmental worldview of children. *International Electronic Journal of Environmental Education*, 2(1), 1–11.
- Carpenter, D. (2014). School culture and leadership of professional learning communities. *International Journal of Educational Management*, 29(5), 682–694.

- Conway, J. A. (1984). The myth, mystery, and mastery of participative decision making in education. *Educational Administration Quarterly*, 20(3), 11–40.
- Creemers, B. (1994). The history, value and purpose of school effectiveness studies. In D. Reynolds, B. P. M. Creemers, et al. (Eds.), *Advances in school effectiveness research and practice* (pp. 9–23). Pergamon, Turkey: Elsevier Ltd.
- Earley, P., & Weindling, D. (2010). *Understanding school leadership*, SAGE Publications: ProQuest eBook Central.
- Eurydice. (2013). *Key data on teachers and school leaders in Europe*. Eurydice Report, Retrieved from EU Publications <https://publications.europa.eu/en/publication-detail/-/publication/17ad39ad-dcad-4155-8650-5c63922f8894/language-en>
- Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of the empirical research. *Educational Administration Quarterly*, 32(1), 27–31.
- Hargreaves, A., & Fink, D. (2009). Distributed leadership: Democracy or delivery? In A. Harris (Ed.), *Distributed leadership – Different perspectives* (Studies in educational leadership series, 7) (pp. 181–196). Dordrecht, The Netherlands: Springer.
- Harris, A. (2009). Distributed Leadership: What we know. In A. Harris (Ed.), *Distributed leadership – Different perspectives* (Studies in educational leadership series, 7) (pp. 101–120). Dordrecht, The Netherlands: Springer.
- Hofman, R. H., Hofman, W. H. A., & Gray, J. M. (2015). Three conjectures about school effectiveness: An exploratory study. *Cogent Education*, 2(1), 1–13.
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.
- Hulpia, H., Devos, G., & Rosseel, Y. (2009a). Development and validation of scores on the distributed leadership inventory. *Educational and Psychological Measurement*, 69(6), 1013–1034.
- Hulpia, H., Devos, G., & Rosseel, Y. (2009b). The relationship between the perception of distributed leadership in secondary schools and teachers' and teacher leaders' job satisfaction and organizational commitment. *School Effectiveness and School Improvement*, 20(3), 291–317.
- IEA. (2017). *Help Manual for the IEA IDB Analyzer* (Version 4.0). Hamburg, Germany. Retrieved from IEA: www.iea.nl/data.html
- Janson, C., Stone, C., & Clark, M. A. (2009). Stretching leadership: A distributed perspective for school counsellor leaders. *Professional School Counselling*, 13(2), 98–106.
- Kennedy, A., Deuel, A., Nelson, T. H., & Slavitt, D. (2011). Requiring collaboration or distributing leadership? *Kappan*, 92(8), 20–24.
- Kruse, S. D., Louis, K. S., & Bryk, A. S. (1995). An emerging framework for analyzing school-based professional community. In K. S. Louis, S. D. Kruse, & Associates (Eds.), *Professionalism and community: Perspectives on reforming urban schools* (pp. 23–44). Thousand Oaks, CA: Corwin Press.
- Lambert, L. (2003). *Leadership capacity for lasting school improvement*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record*, 91(4), 509–536.
- Logan, J. P. (1992). *School-based decision-making: First-year perceptions of Kentucky teachers, principals, and counselors*. Retrieved from EDRS/ERIC Clearinghouse: <https://catalogue.nla.gov.au/Record/5561935>
- Lomos, C. (2017). To what extent do teachers in European countries differ in their professional community practices? *School Effectiveness and School Improvement*, 28(2), 276–291.
- Lomos, C., Hofman, R. H., & Bosker, R. J. (2012). The concept of professional community and its relationship with student performance. In S.G. Huber & F. Algrimm (Eds.), *Kooperation: Aktuelle Forschung zur Kooperation in und zwischen Schulen sowie mit anderen Partnern* (pp. 51–68). Berlin: Waxmann.
- Louis, K. S., Dretzke, B., & Wahlstrom, K. (2010). How does leadership affect student achievement? Results from a national US survey. *School Effectiveness and School Improvement*, 21(3), 315–336.

- Louis, K. S., & Kruse, S. D. (1995). *Professionalism and community: Perspectives on reform in urban schools*. Thousand Oaks, CA: Corwin Press.
- Louis, K. S., & Marks, H. M. (1998). Does professional community affect the classroom? Teachers' work and student experiences in restructuring schools. *American Journal of Education*, *106*, 532–575.
- Louis, K. S., Mayrowetz, D., Smiley, M., & Murphy, J. (2009). The role of sense-making and trust in developing distributed leadership. In A. Harris (Ed.), *Distributed leadership – Different perspectives* (Studies in educational leadership series, 7) (pp. 157–180). Dordrecht, The Netherlands: Springer.
- Melcher, A. J. (1976). Participation: A critical review of research findings. *Human Resource Management*, *15*, 12–21.
- Meuleman, B., & Billiet, J. (2011). Religious involvement: Its relation to values and social attitudes. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 173–206). New York, NY: Routledge, Taylor & Francis Group.
- Meuleman, B., & Billiet, J. (2012). Measuring attitudes toward immigration in Europe: The cross-cultural validity of the ESS immigration scales. *Research and Methods*, *21*(1), 5–29.
- Morrisey, M. (2000). *Professional learning communities: An ongoing exploration*. Austin, TX: Southwest Educational Development Laboratory.
- Muijs, D., West, M., & Ainscow, M. (2010). Why network? Theoretical perspectives on networking. *School Effectiveness and School Improvement*, *21*, 5–26.
- Ning, H. K., Lee, D., & Lee, W. O. (2015). Relationships between teacher value orientations, collegiality, and collaboration in school professional learning communities. *Social Psychology of Education*, *18*, 337–354.
- OECD. (2016). School leadership for developing professional learning communities. *Teaching in Focus 15*(September), 1–4. Retrieved from OECD.ORG - Education Indicators in Focus <http://www.oecd.org/edu/skills-beyond-school/educationindicatorsinfocus.htm>
- Pitner, N. J. (1988). The study of administrator effects and effectiveness. In N. Boyan (Ed.), *Handbook of research in educational administration* (pp. 99–122). New York, NY: Longman.
- Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *Elementary School Journal*, *83*(4), 427–452.
- Rosenholtz, S. J. (1989). *Teachers' workplace: The social organization of schools*. New York, NY: Longman.
- Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 561–582.
- Scherer, R., & Gustafsson, J. E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: An application of multilevel bifactor structural equation modeling. *Frontiers in Psychology*, *6*, 1550.
- Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010). *ICCS 2009 international report: Civic knowledge, attitudes, and engagement among lower-secondary school students in 38 countries*. Amsterdam, The Netherlands: IEA. Retrieved from IEA: https://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ICCS_2009_International_Report.pdf
- Smylie, M. A., Lazarus, V., & Brownlee-Conyers, J. (1996). Instructional outcomes of school-based participative decision-making. *Educational Evaluation and Policy Analysis*, *18*(3), 181–198.
- Spillane, J. P., Halverson, R., & Diamond, J. B. (2004). Towards a theory of leadership practice: A distributed perspective. *Journal of Curriculum Studies*, *36*(1), 3–34.
- Stoll, L., & Louis, K. S. (2007). Professional learning communities: Elaborating new approaches. In L. Stoll & K. S. Louis (Eds.), *Professional learning communities: Divergence, depth and dilemmas* (pp. 1–13). Maidenhead, UK: Open University Press.
- Toole, J. C., & Louis, K. S. (2002). The role of professional learning communities in international education. In K. Leithwood & P. Hallinger (Eds.), *Second international handbook of educational leadership and administration* (pp. 245–279). Dordrecht, The Netherlands: Kluwer.

- Van der Veld, W. M., & Saris, W. E. (2011). Causes of generalized social trust. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 207–248). New York, NY: Routledge, Taylor & Francis Group.
- Wahlstrom, K. L., & Louise, K. S. (2008). How teachers experience principal leadership: The roles of professional community, trust, efficacy, and shared leadership. *Educational Administration Quarterly*, *44*(4), 458–495.
- Wingfield, R. J., Reese, R. F., & West-Olantunji, C. A. (2010). Counselors as leaders in schools. *Florida Journal of Educational Administration and Policy*, *4*(1), 114–130.
- Wohlstetter, P., Smyer, R., & Mohrman, S. A. (1994). New boundaries for school-based management. *Educational Evaluation and Policy Analysis*, *16*, 268–286.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 5

New Ways of Dealing with Lacking Measurement Invariance



Markus Sauerwein and Désirée Theis

5.1 Introduction

Over the past decade, policy-makers have become increasingly interested in studies, such as the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS), in which education systems of various countries are compared. Reforms in education are often based on or legitimated by results of such international studies, and governments may adopt educational practices common in countries that performed well in those studies in an attempt to improve their education system (Panayiotou et al., 2014).

Education can be analyzed at the student, classroom (or teacher), school, and (national) system levels (Creemers & Kyriakidēs, 2008, 2015). Decisions made at the system level (e.g. by policy-makers) affect all other levels. Information about, for example, student achievement or teaching quality in a given country can be compared to that in other countries and used to improve teaching quality. Thus, results of international studies in education, such as PISA, which provides information about students' academic achievement and teaching quality in more than 60 countries, are becoming increasingly interesting to policy makers and might affect classroom processes indirectly through reforms in education, and so on.

However, interpretation of the results of international studies may differ across cultures (Reynolds, 2006). Before a construct (of teaching quality), such as classroom management or disciplinary climate, can be compared across groups (e.g.

M. Sauerwein (✉)

Fließner University of Applied Sciences Düsseldorf, Düsseldorf, Germany
e-mail: sauerwein@dipf.de

D. Theis

DIPF | Leibniz Institute for Research and Information in Education,
Frankfurt am Main, Germany

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_5

countries), the structural stability of that construct needs to be investigated. Thus, measurement invariance (MI) analyses have to be conducted and scalar (factorial) invariance has to be established if mean level changes are to be compared across groups or over time (Borsboom, 2006; Chen, 2007, 2008; van de Schoot, Lugtig, & Hox, 2012).

Until now, MI has been neglected in many studies (e.g. Kyriakides, 2006b; OECD, 2012; Panayiotou et al., 2014; Soh, 2014), which could lead to a false interpretation of the implications of the results. In this paper, we analyze data of the PISA study to explore the effect of lacking MI in studies in which groups are compared. Moreover, we investigate whether lacking MI alone provides information about psychometric properties of the construct under investigation or if it also provides content-related information about the construct. We explore possible explanations for the missing MI by consulting third variables, which are very likely to be equivalent across countries.

5.1.1 The Multi-Level Framework of the Education System

Over the past decade, policy-makers and school administrators have shown an increasing interest in research findings concerning the association between teaching quality and student achievement (Pianta & Hamre, 2009a). Findings of studies, such as PISA, are used to justify and legitimize reforms in education (for a discussion about the influence of PISA findings on policy decisions, see Breakspear, 2012). Accordingly, one goal of studies, such as PISA (OECD, 2010; e.g. OECD Publishing, 2010, 2011) is to identify factors related to students' learning. Some of these factors can be influenced (indirectly) by changes in policy concerning, for example, the curriculum, resource allocation, or teaching quality (e.g. through teacher training or teacher education; Kyriakides, 2006a). The assumption that policy changes affect teaching quality, for example, is based on a multi-level framework of education systems.

The dynamic model of educational effectiveness (Creemers & Kyriakidēs, 2008, 2015; Creemers, Kyriakidēs, & Antoniou, 2013; Panayiotou et al., 2014) describes how system, school, and classroom levels interact. Scheerens (2016, p. 77) states that “within the framework of multi-level education systems, the school level should be seen from the perspective of creating, facilitating and stimulating conditions for effective instruction at the classroom level.” Learning takes place primarily at the classroom level and is associated with teaching quality. At the school level, all stakeholders (teacher, parents, students, etc.) are expected to ensure that time in class is optimized and that teaching quality is improved (Creemers & Kyriakides, 2015). This way, the school level is expected to influence teaching quality (e.g. through regular evaluations at school). The school level, in turn, is influenced by the system/country level through education-related policy, systematic school and/or teacher evaluations, and teacher education (Creemers & Kyriakides, 2015). Hence, policies relevant not only at the classroom level but also at the school and/or country level can improve teaching quality.

5.1.2 Context Matters: Comparing Educational Constructs in Different Contexts

Since the beginning of the twenty-first century, policy-makers have attempted to transfer knowledge and ideas employed in one education system to another (Panayiotou et al., 2014). PISA provides information about students' academic achievement and teaching quality in more than 60 countries. The relation between students' academic achievement and teaching quality is worth being examined at the system level because low scores on achievement tests might correlate with poor teaching quality in a given country. Thus, when students perform poorly on achievement tests, policy-makers might be interested in comparing the teaching quality in their country to the teaching quality in other countries. Detailed knowledge about how students' academic achievement is promoted in various countries might help policy-makers develop appropriate teacher training programs.

As interest in international comparisons in education grows, researchers are becoming increasingly concerned that findings are too simplified and too easily transferred to different cultures (Reynolds, 2006). Comparison of education-related constructs in various subjects, grades, extracurricular activities, and countries requires MI across the different contexts. Hence, to legitimize comparisons of dimensions in different contexts, the dimensions must be stable across the given contexts. MI must be established for the construct under investigation in order to ensure this precondition.

5.1.3 Teaching Quality

Teaching quality often is framed according to the dynamic model of educational effectiveness (Creemers et al., 2013; Creemers & Kyriakidēs, 2008), the classroom assessment scoring system (CLASS) (Hamre & Pianta, 2010; Hamre, Pianta, Mashburn, & Downer, 2007; Pianta & Hamre, 2009a, 2009b), or the three dimensions of classroom process quality (Klieme, Pauli, & Reusser, 2009; Lipowsky et al., 2009; Rakoczy et al., 2007). These models, which show a considerable overlap (Decristan et al., 2015; Praetorius et al., 2018), refer to three essential generic dimensions of teaching quality. The first dimension can be described as classroom management (see also Kounin, 1970) or disciplinary climate. This dimension is closely related to the concept of time on task. It is postulated that clear structures and rules can help students to focus on lessons and to complete tasks (Doyle, 1984, 2006; Evertson & Weinstein, 2006; Kounin, 1970; Oliver, Wehby, & Daniel, 2011). Several studies and meta-analyses have shown a positive correlation between classroom management and students' learning (Hattie, 2009; Kyriakides, Christoforou, & Charalambous, 2013; Seidel & Shavelson, 2007; Wang, Haertel, & Walberg, 1993). The second dimension is cognitive activation or instructional support and refers to (constructivist) learning theories (Fauth, Decristan, Rieser, Klieme, & Büttner,

2014; Klieme et al., 2009; e.g. Lipowsky et al., 2009; Mayer, 2002). The third dimension is commonly referred to as supportive climate, emotional support (e.g. Klieme et al., 2009; Klieme & Rakoczy, 2008), or students' motivation (e.g. Kunter & Trautwein, 2013) and is derived from motivation theories, self-determination theory, in particular (Deci & Ryan, 1985; Ryan & Deci, 2002). In this chapter, we focus on disciplinary climate as a subdimension of classroom management – one central dimension of teaching quality, which is assessed in PISA.

5.1.4 *Measurement Invariance Analyses*

Generally, MI analyses are conducted to determine the psychometric properties of scales and constructs. MI of the construct under investigation across two or more groups or assessment points must be established when (mean) scores of scales, or the influence of a variable on another, are compared because such analyses postulate that the scale measures the same construct in all groups over a certain period of time. If MI is not established, the scale will not measure the same construct in all groups. The results of such comparisons in which MI is not established might be biased and cannot be interpreted as originally intended (Borsboom, 2006; Chen, 2007, 2008; van de Schoot et al., 2012).

MI needs to be distinguished from measurement bias: While bias refers to differences between the estimated parameter and the true parameter, MI refers to comparability across groups (Sass, 2011). Generally, three levels of MI can be differentiated. The most basic level of MI is configural invariance, which is established when items are associated with the same latent construct in different groups or across assessment points. If configural invariance is established, the scale will measure similar but not equal constructs across groups/assessment points. In this case, comparisons of correlations between the scale and other variables in different groups are legitimate. Effect sizes of these correlations, however, should not be interpreted and compared. If configural invariance was not established, scores on the scale under investigation should not be compared across groups or assessment points. The second level of MI is called metric invariance, which is established when factor loadings are equal across groups or assessment points. Value changes in an item for one unit lead to equal changes in the latent construct for all groups. This level of MI allows comparison of associations (and effect sizes) between latent scales and variables across groups or assessment points (Vandenberg & Lance, 2000; Vieluf, Leon, & Carstens, 2010). The third level of MI is scalar invariance, which is established when factor loadings and intercepts of the items representing the latent construct are equal across groups or assessment points. Therefore, the scales share the same intercept. Thus, all groups under investigation have the same starting point, and mean scores can be compared (Chen, 2008; Vandenberg & Lance, 2000).

Recent studies show that the necessary level of measurement invariance for cross-cultural comparisons often is not given (e.g. Vieluf et al., 2010). Moreover,

some studies do not even control for or report MI. Luyten et al. (2005) found that the interactions between socio-economic status (SES) and teaching quality differ across countries, but the authors do not report whether the necessary level of MI (here at least metric MI) for cross-cultural comparisons was established. Similarly, Panayiotou et al. (2014) test the dynamic model of educational effectiveness in different countries and compare the influence of several factors on student achievement, but do not investigate the level of MI for their construct among the different countries (only within the countries) (see also Kyriakides, 2006b and Soh, 2014).

5.1.5 Research Objectives

As mentioned above, results of studies investigating differences in teaching quality across countries are of great interest to policy-makers. Information provided by such studies affects decisions that are made at the system level, which, in turn, affect processes at the classroom level. However, in order to compare certain constructs across groups or over time, invariance among the scales under investigation must be established, which, until now, has not necessarily been the case. The objectives of the present chapter are to

- show how neglecting MI of dimensions under investigation affects results of studies, in which mean levels among groups or assessment points are compared;
- compare the mean score of disciplinary climate among countries;
- investigate whether constructs can be compared even if a certain level of MI is not established; and
- find variables, which could explain the lack of MI among countries.

5.2 Method

5.2.1 Study

We analyzed data from PISA 2009; PISA is a triennial international comparative study of student learning outcomes in reading, mathematics, and science. The focus in PISA 2009 was reading comprehension, which we used as the outcome variable. The reading test in PISA is set at a mean (M) of 500 points and a standard deviation (SD) of 100 points. The study originally was developed as an instrument for OECD countries; now, it is used in more than 65 countries. The study is designed to monitor outcomes over time and provides insights into the factors that may account for differences in students' academic achievement within and among countries (OECD, 2011, 2012).

Students complete a questionnaire assessing, for example, classroom management (measured as disciplinary climate) in the native language lesson (OECD,

2012). Table 5.1 shows the items assessed with this scale (1 = *strongly disagree* – 4 = *strongly agree*) and sample size, means, and the standard deviation of students from Chile, Finland, Germany, and Korea, who participated in PISA 2009. We refer to these countries because they are typical proxies for region-specific educational systems.¹ Furthermore, we use class size as the measurement equivalent variable to explain lacking MI among the countries. For mean and standard deviation of the variable *class size*, see Table 5.2.

Table 5.1 Descriptive statistics of the scale used to assess disciplinary climate in PISA

		Students don't listen to what the teacher says	There is no noise or disorder	The teacher has to wait a long time for students to quiet down	Students cannot work well	Students don't start working for a long time after lessons begin
Chile	<i>M</i>	2.14	2.34	2.22	1.84	2.12
	<i>N</i>	5550	5554	5551	5554	5555
	<i>S.D.</i>	.743	.812	.907	.805	.879
Finland	<i>M</i>	2.40	2.49	2.27	1.94	2.19
	<i>N</i>	5770	5770	5769	5765	5767
	<i>S.D.</i>	.764	.824	.848	.783	.866
Germany	<i>M</i>	1.90	1.86	2.02	1.88	1.84
	<i>N</i>	4420	4430	4424	4390	4417
	<i>S.D.</i>	.780	.830	.871	.838	.888
Korea	<i>M</i>	1.80	2.11	1.72	1.63	1.71
	<i>N</i>	4966	4962	4962	4961	4964
	<i>S.D.</i>	.631	.681	.714	.697	.729
All	<i>M</i>	2.08	2.23	2.07	1.83	1.98
	<i>N</i>	20,706	20,716	20,706	20,670	20,703
	<i>S.D.</i>	.768	.824	.867	.790	.866

M Mean, *S.D.* Standard deviation, *N* Number of students

Table 5.2 Class size

	<i>N</i>	<i>M</i>	<i>SD</i>
Chile	5189	36.16	7.56
Finland	5643	18.77	4.13
Germany	4200	24.66	5.17
Korea	4986	35.98	5.07
All	20,018	28.80	9.51

M Mean, *S.D.* Standard deviation, *N* Number of students

¹Chile represents a South-American system with highly improved rates in PISA tests in the last decades; Germany is well-known for its highly structured education system and is, besides Finland, used as an example for a European system. Korea is a proxy for an Eastern-Asian system with a strong focus on performance and good PISA results. Finland is used as an example for a Scandinavian system, and students are also performing very well in PISA studies.

5.2.2 Data Analyses

Below is a step-by-step explanation of how we compared the scales of the different countries.

1. Comparison of mean levels and associations between disciplinary climate and reading

First, we performed an analysis of variance (ANOVA) to compare mean levels. This allowed us to determine whether there were significant differences in disciplinary climate among the countries. Cohen's d was used to indicate the magnitude of the differences among the countries. Values between .2 and .5 indicated small effect sizes; values between .5 and .8 indicated moderate effect sizes. Higher values (>.8) indicated large effect sizes (Cohen, 1988). Second, we computed regression analyses to identify the association between reading score and disciplinary climate. Including this step before the MI analyses shows how false conclusions can be drawn, if mean levels are compared although MI is lacking. Normally, MI has to be established before mean level scores and effect sizes are compared. However, we turned the normal procedure around in favour of our research objectives.

2. MI analyses and explaining lack of MI

We conducted MI analyses to test the structural stability of the scales used in the context of PISA. A model with parameter constraints was tested against a less restricted model (e.g. metric vs. configural invariance). To determine the level of MI, we compared the fit indices of the models. In line with the literature at hand, we used the comparative fit index (CFI), and the root mean square error of approximation (RMSEA) to test, which model fit the data best (Chen, 2007; Desa, 2014; Sass, 2011; Sass, Schmitt, & Marsh, 2014; Vandenberg & Lance, 2000; Vieluf et al., 2010). A model was accepted, if the fit indices obtained the following scores: CFI > .90, RMSEA < .08 (Hu & Bentler, 1999). In line with results of simulation studies, Chen (2007) recommends that the next higher level of MI be revised, if the CFI decreases by $\geq -.01$ and/or the RMSEA decreases by $\geq .015$. However, Chen (2007, p. 502) states that "[...] these criteria should be used with caution, because testing measurement invariance is a very complex issue." Another way to determine the level of MI is to conduct a chi-square test; however, the results of these tests should be interpreted with caution as they are influenced by sample size. Thus, models designed on the basis of a large sample size could be rejected even if they fit the data well (van de Schoot et al., 2012; Vandenberg & Lance, 2000). The sample studied in PISA is quite large. Thus, we did not conduct chi-square tests. We investigated whether scales or at least single items could be compared among countries. Therefore, we performed the analyses as follows:

- First, we determined the level of MI across all four countries we refer to in our paper (Korea, Finland, Germany, and Chile).
- Second, we determined the level of MI when countries were compared.

- Third, we examined the factor loadings (λ) of the items and investigated whether single items had the same or different (content-related) meaning for the latent construct. To decide, which items had different meanings in different countries, we used the MODINDICES function in MPlus 7.1 (see Muthén & Muthén, 1998–2012). The MODINDICES function provides information about fixed items (between groups) and the expected improvement of model fit if a certain item is freely estimated. Items, which could be fixed between groups, seemed to have the same relevance or meaning for the latent construct in different countries.
- Fourth, we investigated whether single items were comparable. Therefore, we established partial MIs: Some of the factor loadings and/or intercepts among groups were allowed to be estimated freely, while others remained constant (van de Schoot et al., 2013). To decide, which items should be estimated freely, we used again the MODINDICES function in Mplus (Muthén & Muthén, 1998–2012). We allowed factor loadings or intercepts among groups of some items to be estimated freely until the model showed an acceptable fit. This approach allowed us to find items, which were comparable among countries.
- Finally, we tried to identify the reason for possible lacks in MI. We considered variables, which were measurement-invariant by definition among countries. For the purpose of this study, we used the variable class size (see Table 5.2) because a student is a student in every country and therefore comparable across countries.

5.3 Results

5.3.1 *Research Aim No. 1: How Neglecting MI Could Lead to False Interpretations of Results*

Table 5.3 shows the mean levels of the different countries on the scale used to assess disciplinary climate. Without taking MI into account, these results indicate that the highest level of disciplinary climate was reported in Korea. As all differences among the countries are significant ($p < .01$), we also calculated Cohen's d . Our results indicate that there are moderate differences in terms of the mean scores of disciplinary climate between Chile and Korea, Finland and Germany, and Finland and Korea. Moreover, our results show that students in Finland and Korea achieved the highest scores in reading competence (Korea: 539; Finland: 536) (OECD, 2011), but disciplinary climate in both countries differed significantly (Table 5.3). Therefore, we also computed regression analyses to explain the relation between disciplinary climate and reading competence.

As shown in Table 5.4, we found differences in the predictive value of disciplinary climate/classroom management among the countries; in Finland, this effect was very small. Policy-makers in Chile might conclude from these findings that the concept of disciplinary climate in Korea should be adopted in Chile. However, before such conclusions can be drawn, it needs to be tested whether disciplinary climate

Table 5.3 Cohen’s *d* and scores on the reading test

	N	Disciplinary climate Mean	Cohen’s <i>d</i> (differences among the countries)				Reading score – mean
			Chile	Finland	Germany	Korea	
Chile	5567	2.13	–	–0.19	0.35	0.56	449
Finland	5774	2.26	0.19	–	0.53	0.76	536
Germany	4443	1.90	–0.35	–0.53	–	0.17	497
Korea	4972	1.79	–0.56	–0.76	–0.17	–	539

Note: *N* number of students

Table 5.4 Effect of disciplinary climate on reading competences

	<i>B</i>	<i>R</i> ²
Chile	–14.20	0.03
Finland	–6.33	0.01
Germany	–19.48	0.04
Korea	–14.49	0.04

B unstandardized effect of disciplinary climate on Reading Competences (Note: PISA Reading Competence Test has a mean of 500 and a standard deviation of 100)

Table 5.5 MI analyses across countries

	Configural invariance	Metric invariance
CFI	.991	.906
RMSEA	.041	.099

CFI Comparative Fit Index, *RMSEA* Root Mean Square Error of Approximation

has the same meaning in the countries (i.e. Chile and Korea). Therefore, we investigated whether this scale was stable across the different countries, and if mean levels were, thus, comparable.

5.3.2 *Research Aim No. 2: Investigating the Stability of the Scale Used to Assess Disciplinary Climate Across Countries and Comparing Countries Even if MI Is Missing*

First, we determined the level of MI across all four countries. Table 5.5 shows that configural MI was established because there was a meaningful decrease in model fit when we tested the model with greater constraints (metric invariance). This result indicates that mean scores of the latent construct of disciplinary climate cannot be interpreted. The same holds true for the association between this construct and other variables. Thus, it is not legitimate to conclude that the effect of disciplinary climate

on reading competence in Germany is larger than in Finland. In all countries, a similar but not the same construct was measured and solely comparisons of the direction of correlations were legitimate. Hence, one might conclude that there was a positive correlation between students' achievement and disciplinary climate in all countries.

Second, we examined the comparability of countries and ran MI analyses separately for each possible comparison option among the four countries. Table 5.6 illustrates that a comparison of the mean scores between Finland and Chile was legitimate. Here, a better disciplinary climate was reported for Chile ($M = 2.13$) than for Finland ($M = 2.26$). A comparison of the effects of disciplinary climate between Finland and Korea as well as between Chile and Korea was legitimate. In the last case, the model fit (i.e. the CFI and RMSEA) decreased by more than .01. Nonetheless, the fit was acceptable and a comparison might have been legitimate. Thus, here we were able to compare the strength of the relation between disciplinary climate and student achievement.

We found a stronger relation between disciplinary climate and reading competency in Korea than in Finland. In Korea and Chile, the strength of the relation was comparable (see Table 5.4). Comparisons between the other countries were not possible because the necessary level of MI was not established.

Third, we investigated whether the factor loadings of single items in different countries might be interpreted. Table 5.7 shows the factor loadings of the single items. Using the MODINDICES function in MPlus, we were able to conclude from our findings that, for example, items 1 and 2 caused meaningful decreases in the

Table 5.6 Investigating MI among countries

	Configural MI	Metric MI	Scalar MI
Chile – Korea			
CFI	0.990	0.974	.934
RMSEA	0.041	0.054	.075
Chile – Germany			
CFI	0.996	0.927	–
RMSEA	0.028	0.093	–
Germany – Finland			
CFI	0.991	0.904	–
RMSEA	0.042	0.111	–
Chile – Finland			
CFI	.988	.986	.976
RMSEA	.054	.048	.054
Finland – Korea			
CFI	0.985	0.977	0.927
RMSEA	0.053	0.055	0.084
Korea – Germany			
CFI	.994	.880	
RMSEA	.029	.112	

CFI Comparative Fit Index, RMSEA Root Mean Square Error of Approximation, MI Measurement Invariance

Table 5.7 Comparison of factor loadings

	λ factor	<i>S.E.</i>
Chile		
Item 1 – Students don't listen to what the teacher says	0.798	0.015
Item 2 – There is no noise or disorder	0.848	0.011
Item 3 – The teacher has to wait a long time for students to quiet down	0.838	0.010
Item 4 – Students cannot work well	0.815	0.013
Item 5 – Students don't start working for a long time after lessons begin	0.838	0.010
Finland		
Item 1 – Students don't listen to what the teacher says	0.830	0.011
Item 2 – There is no noise or disorder	0.873	0.008
Item 3 – The teacher has to wait a long time for students to quiet down	0.873	0.010
Item 4 – Students cannot work well	0.777	0.017
Item 5 – Students don't start working for a long time after lessons begin	0.825	0.012
Germany		
Item 1 – Students don't listen to what the teacher says	0.914	0.007
Item 2 – There is no noise or disorder	0.955	0.005
Item 3 – The teacher has to wait a long time for students to quiet down	0.944	0.005
Item 4 – Students cannot work well	0.894	0.009
Item 5 – Students don't start working for a long time after lessons begin	0.924	0.006
Korea		
Item 1 – Students don't listen to what the teacher says	0.740	0.028
Item 2 – There is no noise or disorder	0.716	0.027
Item 3 – The teacher has to wait a long time for students to quiet down	0.726	0.025
Item 4 – Students cannot work well	0.858	0.014
Item 5 – Students don't start working for a long time after lessons begin	0.845	0.013

λ factor Factor Loading, *S.E.* Standard Error

model fit (the respective values are not reported on in the table) when Chile and Germany were compared. In the case of Finland and Germany, items 1 and 4 led to a decrease in the model fit. Moreover, items 2 and 3 differed from each other when Korea and Finland were compared. However, here no meaningful decrease in the model fit was found.

Taking Germany and Chile as examples, the MODINDICES in MPlus indicated that fixing the factor loadings of items 1 and 2 led to a decline in model fit. Furthermore, it can be seen in Table 5.6 that the factor loadings for these items differed. To avoid a decline in model fit, we calculated partial metric MI (see van de Schoot et al., 2013). Here, the factor loadings of items 1 and 2 were estimated freely (CFI: .94; RMSEA: .09). Next, we used the MODINCES function again to decide whether more items needed to be estimated freely. However, the analyses produced no model with a satisfying model fit. Thus, mean scores of the scale to assess disciplinary climate in Germany and Chile could not be compared (even if we had merely fixed the factor loading of one item). In the same way, we freely estimated factor loadings between Chile and Korea. Here, the analysis would produce a satisfying

model fit, if we fixed the factor loading of item 4 only (CFI: .99; RMSEA: .04). Hence, a comparison of Chile and Korea for this item (“Students cannot work well”) was justified. Accordingly, we conducted a regression analysis while testing the predictive value of this item in terms of the reading achievement of students in Korea and those in Chile. Results of this analysis indicate that the item had greater predictive value in terms of the Korean students’ achievement in reading than in the reading achievement of the Chilean students (Korea: $B = -16.05$; Chile: $B = -13.93$). Even when the intercept of item four was fixed between Korea and Chile, no meaningful decrease in model fit was found (CFI: .98; RMSEA: .05). Thus, mean scores of this item could be compared between Korea and Chile (Chile: $M = 1.84$; Korea: $M = 1.63$; $p < .01$ | Cohen’s $d = .28$).

Our findings indicate that merely fixing this item led to an acceptable model fit (the factor loadings of all other items were estimated freely). Thus, Chile and Korea can be compared in terms of this single item only even when comparison of single items is seen as critical. Nonetheless, results of the regression analyses indicate that comparing the predictive value of a single item can provide meaningful results. If no comparisons were allowed, however, an interpretation of the different meanings of the items in cultural contexts could be worthwhile. For example, if we wanted to compare Germany and Chile, results of the analysis would indicate that no comparisons are allowed. However, we could say that item 1 (“Students don’t listen to what the teacher says”) is more relevant for the latent construct of disciplinary climate in Germany than in Chile (by comparing factor loadings), and this could be an interesting result on its own.

5.3.3 Research Aim No. 3: Explaining Missing MI by Using Other Variables, Which Are Considered to Have the Same Meaning in Different Countries

Since the meaning of disciplinary climate varied somewhat across the countries under investigation, we searched for possible cultural explanations for the differences in meaning. The challenge here was to find a third variable that definitely had the same meaning in all countries, in other words, a variable, which was measurement-invariant. Thus, if we tried to explain the cultural differences in the meaning of disciplinary climate across the countries by another variable, this variable ought to be culture-invariant so that it can be used as an anchor. One variable that was invariant across the countries under investigation was the number of students in class. This item has the same zero point (=intercept) and the same factor loadings in every country, because a student is counted as one student everywhere and therefore leads to the same decrease of the scale class size. Furthermore, research and practitioners might suggest that classroom size and disciplinary climate are correlated. Thus, we used the number of students in class as an anchor

when trying to explain the cultural differences in the concept of disciplinary climate. We conducted several regression analyses: We used the entire scale as a dependent variable and five single items related to disciplinary climate as dependent variables. In all models, the number of students was used as the independent variable. We conducted these analyses separately for Chile, Finland, Korea, and Germany.

In Chile and Finland, the number of students in class predicted disciplinary climate (see Table 5.8). In these countries, disciplinary climate became more problematic as the number of students in class increased. We found the opposite effect in Korea: A large number of students in class correlated positively with disciplinary climate. In Finland and Chile, the number of students in class also correlated with items 2, 3, and 5. In Korea, the opposite effect was found when item 2 was used as the outcome variable. For Germany, we found no effects.

In summary, our results indicate that the number of students in class can be used as a variable to explain why disciplinary climate has the same meaning (scalar) in Chile and Finland and why, thus, mean levels are comparable in these countries. In these countries, disciplinary climate is associated with the same invariant third variable, and this might – but not must – be a reason why we find scalar MI between Chile and Finland. Furthermore, we found that comparisons of mean scores or correlations between disciplinary climate and other variables (e.g. reading comprehension) were not legitimate between Germany and other countries. Here, class size had no effect on disciplinary climate, which supports our interpretation described above. In Korea, the effects of number of students in class were inversed to Finland and Chile but still had predictive value. This might be the reason why disciplinary climate had a similar meaning in these countries (metric MI) but not the same meaning, which allows mean score comparisons; mean level comparisons were not allowed. However, we can compare the relation between disciplinary climate and reading competencies in Korea with that in Chile, and Finland.

Table 5.8 Regression analysis: independent variable = number of students in class; dependent variable = scale of disciplinary climate as well as the single items of scale separately

	Chile <i>B</i>	Finland <i>B</i>	Korea <i>B</i>	Germany <i>B</i>
Disciplinary Climate	.030*	.074***	-.053***	-.012
Item 1 – Students don't listen to what the teacher says	.019	.103***	-.018	.020
Item 2 – There is no noise or disorder	.030*	.074***	-.053***	-.012
Item 3 – The teacher has to wait a long time for students to quiet down	.042**	.088***	.019	.012
Item 4 – Students cannot work well	.007	.012	-.024	.002
Item 5 – Students don't start working for a long time after lessons begin	.027*	.041***	-.022	.027

Note: * = $p < .05$, ** = $p < .01$, *** = $p < .001$

5.4 Discussion

Our results underline the importance of MI analyses in international comparative educational studies. Analyses based on PISA 2009 data show that results of such studies might be biased or misinterpreted, if MI was not tested before any further analyses are conducted. However, our findings also suggest that more detailed analyses would be worthwhile.

If MI was ignored, our findings indicated that students in Finland and Korea achieve high scores in terms of reading achievement while the mean level of disciplinary climate differed significantly between these countries. Moreover, the predictive value of disciplinary climate for the students' reading achievement differed significantly between these countries as well. Especially in Finland, the effect of disciplinary climate on reading achievement was rather low. The finding that classroom management (disciplinary climate) was an important predictor for students' learning is in line with findings from earlier studies (Carroll, 1963; Seidel & Shavelson, 2007). Such findings might be particularly valuable to policy-makers. For example, policy-makers in Germany might conclude that in good education systems, like the one in Finland, disciplinary climate is not relevant for student achievement. As a result, disciplinary climate might not be included as an indicator of teaching quality in schools or teacher evaluations anymore. However, these findings need to be treated with caution as they stem from analyses that are not legitimate from a methodological point of view. Analyses and interpretations, as they were described in this section, postulate that the constructs under investigation have the same meaning across groups. MI analyses, however, indicate that only configural MI was established in the scales we used; thus, mean levels in the different countries cannot be compared. Nonetheless, we recommend further analyses to be conducted in which findings from different countries will be compared. Additionally, our findings indicate that analyzing levels of MI based on single items can be worthwhile: In Chile – for the factor disciplinary climate – it is important to be quiet during lessons (item 2), and that teachers do not have to wait too long until lessons can start (item 3). If Germany and Chile were compared, it seemed that in Germany, the first item (“Student’s don’t listen to the teacher”) as well as the second item (“There is no noise or disorder”) were more relevant for the disciplinary climate. Comparing Finland and Germany showed that in Finland, item 1 (“Students don’t listen to what the teacher says”) and item 4 (“Can’t work well”) were not as meaningful as they were in Germany. The interpretation of factor loadings as a result on its own seems to be uncommon. However, this idea is similar to interpretations of differential item functioning (DIF) in the context of test construction and scaling (Klieme & Baumert, 2001; see also Greiff & Scherer, 2018). One possible explanation for differences in factor loadings could be that students in different countries/cultures have a different system of relevance for disciplinary climate, and therefore the meaning of disciplinary climate differs among countries/cultures. Teaching and behaviour during class are liable to cultural contexts. This is also underlined by different factor loadings.

If a construct compared between two groups does not meet the standards of MI, the construct conceptually conveys different meanings in these groups (Chen, 2008). Creemers and Kyriakides (2009), for example, report that the development of a school policy for teaching and evaluation has stronger effects in schools where the quality of teaching at the classroom level is low. However, this conclusion could be drawn only if a necessary level of MI was established, otherwise the conclusion drawn may be wrong. If research on school improvement and school effectiveness aimed to compare models in different countries – such as the dynamic model of educational effectiveness – the level of MI should be investigated and proven as a precondition of further analyses. A good example of how to determine and deal with MI in international studies has been described in a very detailed technical report of the TALIS study (OECD, 2014; Vieluf et al., 2010). Moreover, even if MI is missing for the entire scale, it is possible to identify single countries or items for comparison. As a preliminary step, not a multi-group CFA should be conducted with all countries in one model, but rather single countries should be selected for comparison. This might help researchers identify several countries for comparison. If scalar invariance is not given in the countries under investigation, it would be possible to identify single items that can be compared in a next step.

The analyses presented in this paper show that missing MI is not a reason for desisting from comparisons (between pedagogical contexts or cultures). Our findings indicate that the meaning of disciplinary climate differs among cultural contexts. In our opinion, this result should also be reported as a result of its own (see also Greiff & Scherer, 2018, for that issue). Given the fact that research in education is used as a tool to legitimate policy actions and that results are transferred from one cultural context to another, reporting missing MI appears to be especially important (Martens & Niemann, 2013; Panayiotou et al., 2014; Reynolds, 2006). Even if schools within a country were compared, MI should be tested because all schools differ from one another and might have their own school culture. Therefore, conclusions that the development of a school policy for teaching and external evaluation have been found to be more influential in schools where the quality of teaching at the classroom level is low (Creemers & Kyriakides, 2009) should be treated with caution.

Furthermore, qualitative methods (e.g. documentary methods, such as comparative analyses of different milieus, fields, cultural experiences, etc.; Bohnsack, 1991) refer to different systems of relevance people have, due to different structures of everyday life. The aim of this method is not to compare certain manifestations or means but rather to explain differences. This methodological background can be used to interpret the result of missing MI. In the case of lessons, we can assume that students have different systems of relevance when they are rating classroom management or disciplinary climate. In other words, students do not refer to the same standards when they rate lessons. Thus, we have good reasons to interpret missing MI as an important result. Theoretically, this reasoning is also in line with Lewin's field theory (Lewin, 1964). Person, context, and environment influence and depend on each other. Hence, teaching quality is nested in its cultural and pedagogical context. "Teachers' work does not exist in a vacuum but is embedded in social, cultural,

and organizational contexts” (Samuelsson & Lindblad, 2015, p. 169). A high-quality teacher in India does not allow questioning by students whereas in classes in the United States of America, the opposite is true (Berliner, 2005). Differences in factor loadings and intercepts could be seen as an expression of the cultural and institutional varieties, which should be considered more in international comparative studies. Furthermore, new possibilities may present themselves to identify what cultures display similar facets of teaching, schools, and the education system and therefore what characteristics thereof could be transferred to other education systems.

5.5 Conclusion

This paper presents one of the first attempts to interpret (lacking) MI not only from a methodological point of view but also in terms of content. Chen (2008) explains missing MI for the construct *self-esteem* between China and the USA. Our results indicate that the lack of MI can be seen as a result as well. Nevertheless, we propose further analyses that might investigate ways to compare at least parts of constructs. In summary, our approach to interpreting MI is in line with those of many researchers investigating school improvement and school development, who emphasize the local context of schools and stress the importance of international comparisons (Hallinger, 2003; Harris, Adams, Jones, & Muniandy, 2015; e.g. Reynolds, 2006). The analyses presented here make it possible to identify comparable single cross-cultural items.

References

- Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, 56(3), 205–213. <https://doi.org/10.1177/0022487105275904>
- Bohnsack, R. (1991). *Rekonstruktive Sozialforschung. Einführung in Methodologie und Praxis qualitativer Forschung*. Opladen: Leske + Budrich.
- Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, 44(Suppl 3), S176–S181. <https://doi.org/10.1097/01.mlr.0000245143.08679.cc>
- Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance* (OECD Publishing, Ed.) (OECD Education Working Papers no. 71).
- Carroll, J. (1963). A model of school learning. *Teachers College Record*, 64, 723–733.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464–504.
- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95(5), 1005–1018. <https://doi.org/10.1037/a0013193>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale: L. Erlbaum Associates.
- Creemers, B., & Kyriakides, L. (2009). Situational effects of the school factors included in the dynamic model of educational effectiveness. *South African Journal of Education*, 29, 293–315.

- Creemers, B., & Kyriakides, L. (2015). Developing, testing, and using theoretical models for promoting quality in education. *School Effectiveness and School Improvement*, 26(1), 102–119. <https://doi.org/10.1080/09243453.2013.869233>
- Creemers, B. P. M., & Kyriakides, L. (2008). The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools. In *Contexts of learning*. London, UK/New York, NY: Routledge.
- Creemers, B. P. M., Kyriakides, L., & Antoniou, P. (2013). *Teacher professional development for improving quality of teaching*. Dordrecht, The Netherlands/New York, NY: Springer.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior: Perspectives in social psychology*. New York, NY: Plenum.
- Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Buttner, G., Fauth, B., ... Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal*, 52(6), 1133–1159. <https://doi.org/10.3102/0002831215596412>
- Desa, D. (2014). *Evaluating measurement invariance of TALIS 2013 complex scales: Comparison between continuous and categorical multiple-group confirmatory factor analyses* (OECD education working papers, no. 103). Retrieved from OECD website: <https://doi.org/10.1787/5jz2kbbvlb7k-en>
- Doyle, W. (1984). How order is achieved in classrooms: An interim report. *Journal of Curriculum Studies*, 16(3), 259–277. <https://doi.org/10.1080/0022027840160305>
- Doyle, W. (2006). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management. Research, practice, and contemporary issues* (pp. 97–125). Mahwah, NJ: Lawrence Erlbaum Associates.
- Evertson, C. M., & Weinstein, C. S. (Eds.). (2006). *Handbook of classroom management: Research, practice, and contemporary issues*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>
- Greiff, S., & Scherer, R. (2018). Still comparing apples with oranges? *European Journal of Psychological Assessment*, 34(3), 141–144.
- Hallinger, P. (2003). Leading Educational Change: reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education*, 33(3), 329–351.
- Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 25–41). New York, NY: Routledge.
- Hamre, B. K., Pianta, R. C., Mashburn, A., & Downer, J. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms* (Foundation for Child Development, Ed.).
- Harris, A., Adams, D., Jones, M. S., & Muniandy, V. (2015). System effectiveness and improvement: The importance of theory and context. *School Effectiveness and School Improvement*, 26(1), 1–3. <https://doi.org/10.1080/09243453.2014.987980>
- Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement*. London, UK: Routledge.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education*, 16(3), 385–402. <https://doi.org/10.1007/BF03173189>
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster, Germany: Waxmann.

- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcomeorientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54(2), 222–237. Accessed 29 Dec 2012. https://www.pedocs.de/frontdoor.php?source_opus=4348
- Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York, NY: Holt Rinehart & Winston.
- Kunter, M., & Trautwein, U. (Eds.). (2013). *Standard Wissen Lehramt: Vol. 3895. Psychologie des Unterrichts*. Paderborn, Germany: Schöningh.
- Kyriakides, L. (2006a). Introduction international studies on educational effectiveness. *Educational Research and Evaluation*, 12(6), 489–497. <https://doi.org/10.1080/13803610600873960>
- Kyriakides, L. (2006b). Using international comparative studies to develop the theoretical framework of educational effectiveness research: A secondary analysis of TIMSS 1999 data. *Educational Research and Evaluation*, 12(6), 513–534. <https://doi.org/10.1080/13803610600873986>
- Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, 36, 143–152.
- Lewin, K. (1964). *Field theory in social science*. New York, NY: Harper & Brothers.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6), 527–537. <https://doi.org/10.1016/j.learninstruc.2008.11.001>
- Luyten, J. W., Scheerens, J., Visscher, A. J., Maslowski, R., Witziers, B. U., & Steen, R. (2005). *School Factors related to quality and equity. Results from PISA 2000* (OECD, Ed.).
- Martens, K., & Niemann, D. (2013). When do numbers count?: The differential impact of the PISA rating and ranking on education policy in Germany and the US. *German Politics*, 22(3), 314–332. <https://doi.org/10.1080/09644008.2013.794455>
- Mayer, R. E. (2002). Understanding conceptual change: A commentary. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change. Issues in theory and practice* (pp. 101–111). Dordrecht, The Netherlands: Kluwer Academic.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus statistical analysis with latent variables: User's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- OECD. (2010). *PISA 2009 results*. Paris, France: OECD.
- OECD. (2011). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science (volume I) (1. Aufl.)*. PISA 2009 results. s.l. Paris, France: OECD.
- OECD. (2012). *PISA 2009 technical report*. Paris, France: OECD.
- OECD. (2014). *TALIS 2013 technical report*.
- Oliver, R., Wehby, J., & Daniel, J. (2011). *Teacher classroom management practices: Effects on disruptive or aggressive student behavior*. Campbell Systematic Reviews 2011.4.
- Panayiotou, A., Kyriakides, L., Creemers, B. P. M., McMahon, L., Vanlaar, G., Pfeifer, M., ... Bren, M. (2014). Teacher behavior and student outcomes: Results of a European study. *Educational Assessment, Evaluation and Accountability*, 26(1), 73–93. <https://doi.org/10.1007/s11092-013-9182-x>
- Pianta, R. C., & Hamre, B. K. (2009a). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. <https://doi.org/10.3102/0013189X09332374>
- Pianta, R. C., & Hamre, B. K. (2009b). Classroom processes and positive youth development: Conceptualizing, measuring, and improving the capacity of interactions between teachers and students. *New Directions for Youth Development*, 2009(121), 33–46. <https://doi.org/10.1002/yd.295>
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality. The German framework of Three Basic Dimensions. *ZDM*, 50, (3, 407–426). <https://doi.org/10.1007/s11858-018-0918-4>
- Publishing, O. (2010). *PISA 2009 Results: Learning to Learn: Student Engagement, Strategies and Practices (Volume III) (1. Aufl.)*. PISA 2009 Results. s.l. Paris, France: OECD.

- Publishing, O. (2011). *PISA quality time for students: Learning in and out of school (1. Aufl.)*. PISA. s.l. Paris, France: OECD.
- Rakoczy, K., Klieme, E., Drollinger-Vetter, B., Lipowsky, F., Pauli, C., & Reusser, K. (2007). Structure as a quality feature in mathematics instruction: Cognitive and motivational effects of a structured organisation of the learning environment vs a structured presentation of learning content. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG priority programme* (pp. 101–120). Münster, Germany/New York, NY/München, Germany/Berlin, Germany: Waxmann.
- Reynolds, D. (2006). World Class Schools: Some methodological and substantive findings and implications of the International School Effectiveness Research Project (ISERP). *Educational Research and Evaluation*, 12(6), 535–560. <https://doi.org/10.1080/13803610600874026>
- Ryan, R., & Deci, E. (2002). An overview of self-determination theory: An organismic-dialectical perspective. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3–33). Rochester, NY: University of Rochester Press.
- Samuelsson, K., & Lindblad, S. (2015). School management, cultures of teaching and student outcomes: Comparing the cases of Finland and Sweden. *Teaching and Teacher Education*, 49, 168–177. <https://doi.org/10.1016/j.tate.2015.02.014>
- Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 29(4), 347–363. <https://doi.org/10.1177/0734282911406661>
- Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 167–180. <https://doi.org/10.1080/10705511.2014.882658>
- Scheerens, J. (2016). *educational effectiveness and ineffectiveness: A critical review of the knowledge base* (1st ed.). Dordrecht, The Netherlands: Springer.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. <https://doi.org/10.3102/0034654307310317>
- Soh, K. (2014). Finland and Singapore in PISA 2009: Similarities and differences in achievements and school management. *Compare: A Journal of Comparative and International Education*, 44(3), 455–471. <https://doi.org/10.1080/03057925.2013.787286>
- van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00770>
- van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486–492. <https://doi.org/10.1080/17405629.2012.686740>
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70.
- Vieluf, S., Leon, J., & Carstens, R. (2010). Construction and validation of scales and indices. In *TALIS 2008 technical report* (pp. 131–206).
- Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, 63(3), 249–294. <https://doi.org/10.3102/00346543063003249>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 6

Taking Composition and Similarity Effects into Account: Theoretical and Methodological Suggestions for Analyses of Nested School Data in School Improvement Research



Kai Schudel and Katharina Maag Merki

6.1 Expanding the Concept of Group Level in School Research

Increasingly, theoretical and empirical studies have shown that the teaching staff plays an important role in school improvement and in fostering student learning, since regulations, guidelines, and the decisions on the system level and on the level of the school management (school leader) have to be re-contextualized by the teaching staff and individual teachers to exert their influence on student learning and student outcomes (Fend, 2005, 2008; Hallinger & Heck, 1998). To deal with such processes, multilevel analysis has proven to be the standard in empirical school research (Luyten & Sammons, 2010). In this contribution, the multilevel approach is expanded to include a theoretical and methodological focus on the double character of group levels in organizations, on composition effects on a group level, and on position effects on an individual level.

Multilevel models allow depiction of hierarchically structured phenomena, such as schools or classes. For example, separate students are gathered in a single classroom, which is often assigned to a specific teacher. Separate teachers, in turn, form a teaching staff and a school, and separate schools are administrated by a school board in a municipality. Finally, schools are part of a geographical entity.

Analysing this nested or clustered structure as a multilevel model is a methodological necessity for two reasons. First, it considers the fact that observations of the same unit are not independent. Thus, it counteracts overestimation of statistical findings, as observations that belong to the same unit on a higher level are interdependent. It also allows determination of the contribution of the different

K. Schudel (✉) · K. Maag Merki
University of Zurich, Zurich, Switzerland
e-mail: kai.schudel@ife.uzh.ch

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_6

levels regarding the overall variance of an interesting feature on the lowest level (Luyten & Sammons, 2010). Therefore, differences in student achievement, for example, can be attributed in a more differentiated manner to influences of the separate students, teachers, school management, the school, and possibly also to city districts.

But the way that nested structures are usually considered and calculated by multilevel models indicates a limited understanding of what non-independence of observations within a unit or a group means. This becomes clear by the fact that measures of agreement, such as the intraclass correlation (ICC), is usually used to determine the necessity of a multilevel model. Intraclass correlation (ICC) represents the ratio of the variance between units to the total variance, and it is interpreted as a measurement of agreement or similarity among observations within a unit (LeBreton & Senter, 2007). Therefore, when non-independence is conceived of only as the presence of a significant ICC value, the non-independence is simply defined by an over-proportional similarity of observations within a unit. But non-independence can mean more than converging observations, such as, for example, same shared attitudes among teachers or the same teaching staff. Non-independence in nested structures can be defined more generally by simply acknowledging that observations are influenced by the unit that they are in, and thus, by the shared context, and the unit's influence can manifest itself in various forms. For teachers on a teaching staff, for example, the shared unit does not have to lead to shared attitudes. The same shared unit can also result in different attitudes because the teaching staff serves as an umbrella under which teachers have to interact. In this sense, non-independence means that every teacher refers to the other teachers within the same teaching staff. Thus, each teaching staff can be described by a specific composition and pattern that are a result of non-independence of the teachers.

This problem of too simplified group-level conceptions and non-independence has also been criticized in research on small groups and in organizational research by Kozlowski and Klein (2000). They also point out that research often simply aggregates lower-level individual characteristics to the next higher group level by averaging, without considering that groups can also be described by the specific composition of the individual characteristics. They suggest that groups and, thus, every higher level in nested data can be described by *global properties*, *shared properties*, and *configural properties*. We can adopt these aspects in our criticism of school research above. Global properties are located at the group level, or the higher level, respectively; they manifest only on that level, and their measurement does not depend on lower-level characteristics and are thus non-controversial. Therefore, global properties of a group serve as a shared context for lower level individuals. Furthermore, because they serve as a context for the individuals on lower level, global properties initiate a top-down process (Kozlowski, 2012). Collective characteristics of the lower level, which describe how similar or dissimilar group members are, can be generally described by *group composition* (Kozlowski, 2012; Lau & Murnighan, 1998; Mathieu, Maynard, Rapp, & Gilson, 2008; Schudel, 2012). According to Kozlowski and Klein (2000), the composition of a group can be described by shared properties or by configural properties. Shared properties are

those characteristics of individuals that converge within the group and represent the homogeneity thereof. Configural properties are those characteristics of individuals that diverge within the group and represent the heterogeneity of a group.

In the case of school research, the neglect of group composition may be connected to the *double character* that group levels in school environment usually possess. The entities on a higher level – such as schools or classrooms – can be described by either separate characteristics on that higher level – the global properties – or by collective characteristics on a lower level – the group composition. Global properties can be an area of responsibility of a single individual on the higher level or a shared higher-level context. However, collective characteristics on a group level can only be described by the interplay of multiple individuals on the lower subordinate level. They emerge from the lower level by interaction but manifest themselves at the group level; thus, group composition refers to the fact that what develops in a group is more than just the simple sum of the individuals (Kozłowski & Klein, 2000). Therefore, the information about the global properties of a group can be obtained from that group level, and the information about group composition can only be gathered from the multiple lower level entities. For instance, if we are interested in the school level, we can describe and measure the global properties by separate characteristics of the responsible school principal or of the school, such as leadership quality and budget. But we can also describe and measure the composition of the school by collective characteristics of the cluster of teachers working at the school, the shared and configural properties of the teaching staff, such as shared beliefs of the teachers, but also as diverging subjective perspectives. The same holds true for the classroom level: We can describe and measure the global properties by separate characteristics of the responsible class teacher or of the classroom infrastructure, such as teaching quality and the number of computers available. We can also describe and measure the classroom composition by collective characteristics of the cluster of students that form a class, e.g. the average school achievement of the students as a shared property, when we assume that students in a class tend to have a similar learning progress – or e.g. different educational family backgrounds as a configural property.

In conclusion, although multilevel models in school research acknowledge that a group level always constitutes a combination of entities of a lower level (e.g. teaching staff as an association of teachers), the underlying assumption usually is that the shared group context leads to homogeneous entities. Therefore, research often focuses solely on shared properties, which is represented by the calculation of a group mean. However, the explanations above show that non-independence and shared group context do not preclude the possibility that the lower-level entities or individuals are different. Therefore, multilevel models in school research have to consider the double character of groups, consisting of global group properties emerging from the group level, and group composition emerging from the individual lower level. Further, they have to consider the possibility of both shared properties and configural properties of group compositions.

Disentangling those two characteristics of a group or a higher level entity is also crucial because it allows us to depict the re-contextualization processes in the school

environment (Fend, 2005, 2008). If we separated global properties from group composition, we could make it visible that global properties – such as a responsible person or an existing infrastructure – serve as an opportunity and that individuals on the lower level make use of that opportunity by their specific group composition. Kozlowski (2012) analogously observes that a group is finally the result of top-down effects of global properties and bottom-up effects emerging from the group composition. That what we measure on a specific unit level, therefore, is mostly a result of the interactions between a responsible separate person, or a shared context characteristic, and a subordinate collective as shown in Fig. 6.1.

As composition and configural properties in particular are often missing in research, we can assume that research reduces unit levels to areas of responsibility rather than also take their collective character of associations into account. Therefore, contrary to the theoretically acknowledged fact that diversity of the teaching staff has an influence on school improvement processes, research has placed too little emphasis on the compositional characteristics and composition effects of the teaching staff in study designs and analyses.

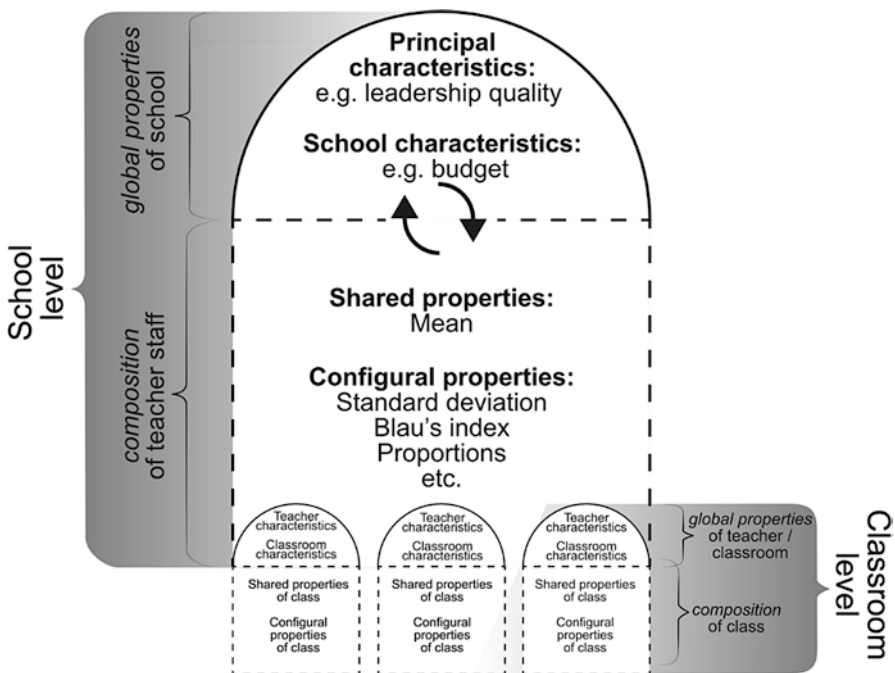


Fig. 6.1 Double character of group levels in school research
 Group levels can be described by separate global properties (semi-circles) and by collective composition (dashed rectangles). Group compositions emerge from subordinate lower level entities and can be described by shared properties and by configural properties. A group is a product of top-down effects of global properties and bottom-up effects of group composition

At class level, the well-known ‘little-fish-big-pond effect’ can be taken as an example: A student’s self-concept is affected not only by his or her own achievements, but also by the aggregated average performance index of the classroom (the entity one level above the student). Accordingly, the school class acts as a frame of reference, through social comparison, for students’ self-concepts (Marsh et al., 2008). This is a phenomenon at the classroom level, and it has also been understood as a composition effect.

Further, pertaining to the level of the teachers, the literature on school improvement capacity or professional learning communities points to the importance of group composition. Mitchell and Sackney (2000), for example, emphasize the relevance of interpersonal capacities to learning communities. This relevance becomes apparent in shared properties, such as shared norms, expectations, and knowledge, or in communication patterns, among other things. For group climate to be effective, each group member’s contributions should be explicitly acknowledged. As a consequence, Mitchell and Sackney (2000) also observed problems in schools with high configural properties, thus, with group compositions, in which dominant excluding subgroups were formed that isolated and marginalized other members. Also, Louis, Marks, and Kruse (1996) showed that diverse subgroups within the teaching staff can have negative effects on the successful achievement of joint objectives. They assume that subgroups can emerge particularly in large schools, alongside discipline demarcations. However, despite the relevance of the composition and structure of teaching staff, there are (still) no studies examining these composition effects differentially.

Based on diversity research, we will first elaborate on how composition can be theorized in school improvement research, particularly at the teaching staff level. In a second step, the Group Actor-Partner Interdependence Model (GAPIM) approach is introduced as a methodological tool. The GAPIM allows analysis of composition effects on the individual level and takes the particular position of the teachers on staff into consideration. We then apply the model to an existing data set (Maag Merki, 2012) as an example.¹ We will illustrate the analysis of the main effects and composition effects of the teaching staff and positioning effects of the separate teachers on the teaching staff regarding the effects of teachers’ individual and collective self-efficacy on teachers’ individual job satisfaction. Since in the existing study, teachers at 37 secondary schools completed a standardized survey on various aspects, the data set is suitable to discuss strengths and weaknesses of the GAPIM for school improvement research.

¹Originally, Maag Merki (2012) analyzed the effects of the implementation of state-wide exit examinations on school, teachers, and students in 37 German upper secondary schools (ISCED 3a). The present contribution, however, does not focus on the analyses of the effects of the implementation of state-wide exit examinations.

6.2 Composition Effect as Diversity Typologies

As mentioned above, the composition of a group can be described by converging or diverging characteristics represented by shared and configural properties. In order to conceptualize different types of shared and configural properties, approaches from diversity research and particularly the typology of Harrison and Klein (2007) are useful (Schudel, 2012).

Diversity of teams is of great importance in the concept of learning communities and distributed leadership (Hargreaves & Shirley, 2009; Mitchell & Sackney, 2000; Stoll, 2009). But diversity can have diverging consequences. It can lead to lower levels of communication through social categorization processes, but [at the same time] it can lead to higher levels of problem solving when diversity reflects a variety of different qualities (Van Knippenberg, de Dreu, & Homan, 2004; Van Knippenberg & Schippers, 2006). This twofold character of diversity is a central issue in research on small groups and is discussed theoretically from an interference-oriented perspective and a resource-oriented perspective (Schudel, 2012). In the context of school improvement, Mitchell and Sackney (2000) point out that diversity endangers a teaching staff, if it leads to the formation of subgroups and, in doing so, undermines shared norms and cooperation. In contrast, the potential of diversity is expressed in the demand “to make a cultural transformation so as to embrace diversity rather than to demand homogeneity” (Mitchell & Sackney, 2000, p. 14). A more differentiated theoretical account of diversity is needed in order to account for the composition effects of teams.

Harrison and Klein (2007) differentiated three types of diversity: *separation*, *variety*, and *disparity*. This differentiation provides a basis for both the interference-oriented perspective and the resource-oriented perspective. With *separation*, diversity can be described as a measure for the formation of subgroups. It is based on similarities between group members regarding a distinct feature, a position or opinion, quantified along a continuum. Consequently, teachers can be compared with each other, for example regarding their tenure – i.e. their position along the continuous attribute tenure. Separation describes the level of similarity between group members. This level is expressed statistically through the standard variation of the feature on the group level. Therefore, a teaching staff exhibits a high level of separation, if the teachers hold positions on both extreme poles of the specific feature’s continuum, such as when half of the teachers have only recently been employed at the school while the other half have been working there for a long time. There is a moderate degree of separation when the teachers are distributed evenly over the continuum of the feature. There is a small degree of separation when all teachers hold the same position on the continuum of the feature, such as when they all have been employed at the school for an equally long time. Since separation is a symmetrical similarity measure, it would be irrelevant at a low level of separation, if all teachers exhibited a long or a short term of employment. Relevant would only be that they exhibited a similarly long or similarly short term of employment. Therefore, separation constitutes a conceptualization in accordance with the practically

relevant potential of subgroup formation within a teaching staff. From an interference-oriented perspective, high separation would have negative consequences for communication and interaction.

The second type of diversity, following Harrison and Klein (2007), is *variety*. The term variety describes the presence of different resources and qualities within a group. It is based on different features of group members that are not quantitatively comparable on a continuum but are of different qualities. For example, teachers are able to form a more or less diverse and heterogeneous teaching staff regarding their subject(s), function, or discipline. Therefore, variety describes the heterogeneity of categorically different features or qualities. Statistically, this is expressed in Blau's index (1977), describing the number of different categories available within a group. Therefore, the teaching staff possesses the highest variety, if all members of the teaching staff teach a different subject, for example. There would be minimal variety in this respect, if all teachers taught the same subject, or, in other words, if the school was highly specialized. Variety is thus operationalized as the different qualitative backgrounds of the teaching staff. It reflects the presence of different kinds of knowledge and abilities in the sense of informational diversity. From a resource-oriented perspective, high variety could therefore be beneficial for problem-solving in community learning (Jehn, Northcraft, & Neale, 1999). Yet, from an interference-oriented perspective, high variety could also describe potential difficulties for divided norms and values and commitment in big and fully differentiated schools (Louis et al., 1996).

Finally, as a third type of diversity, *disparity* means the distribution of hierarchically structured resources within a group. It is based on the distribution of certain normatively desired or valuable features within a group – such as power, wealth, status, or privileges – that are understood as scarce resources. Disparity is, therefore, an asymmetrical measure. It makes a difference whether a minority or a majority holds most of the resources. For example, teaching staffs can differ in how competencies and decisional power are equally distributed among the teachers. Statistically, disparity is expressed in the proportional relation between group members and resource allocation. The teaching staff exhibits a high level of disparity, if, for example, a minority of teachers possess the most – or an unproportioned amount of – decisional power. A lower level of disparity prevails, if the teaching staff has a flat hierarchy, and all teachers have a similar amount of decision-making authority. Disparity is thus able to describe, for example, how much say the teachers have in important decisions and how strongly they are included/involved in the development of changes. Disparity can offer an important indicator of the distributed leadership status (Stoll, 2009).

The three diversity types describe the composition of groups. Instead of reducing the teaching staff to its shared properties and solely considering its group means, school improvement research has to take the multi-faceted composition of the teaching staff into account. Furthermore, Harrison and Klein's (2007) diversity typology not only reveals additional important *descriptive* information about characteristics of shared and configural properties of the teaching staff, but can also be used in causal analyses. The composition measures of the teaching staff can be modelled as

results of antecedent processes. Good school leadership, for example, can result in a teaching staff with low separation, high variety, and low disparity. Or, alternatively, the composition measures of the teaching staff can be modelled as causes of the outcomes of schools, teaching staffs, and separate teachers. For example, from an interference-oriented perspective, high separation of a teaching staff can result in low performance of the school, in low cooperation within the teaching staff, and in low job satisfaction in separate teachers. As a result, these measures introduce new insights into school development research regarding how the teaching staff is structured, what causes this structure, and to what extent the structure has an influence on teacher outcomes, the development of curricula, or the learning curve of students.

6.3 Positioning Effect

Now, if group compositions of this kind are to be examined as predictors of dependent variables on a subordinate individual level, the three diversity types by Harrison and Klein (2007), presented above, have theoretical and methodological shortcomings. Further considerations are necessary that incorporate the individual level.

Diversity, conceptualized on only the group level, abstracts from the *definite position* of the single individual within the group. However, if group composition is taken as a predictor of effects on the individual level, this definite position of the individual within the group composition will not be ignored. Accordingly, group composition signifies different things, depending on the position of a person within this diversity. Naturally, this is most evident in the asymmetrical group composition of disparity. For example, depending on where teachers are within a group characterised by a high level of disparity, they are in possession of resources or not. But also regarding symmetrical measures, such as separation and variety, there are differences in teachers' positions within the compositions of their groups. For example, a group might exhibit a low level of separation or variety. Yet, if a single teacher deviated from such an otherwise homogeneous group, that person could perceive their individual position as isolated. A moderate separation of the teaching staff regarding tenure can have different effects for those teachers that exhibit average tenure (and, thus, are positioned along the continuum in the middle) as compared to newly employed teachers and the most senior teachers (and, thus, those positioned at one of the extreme poles).

Kenny and Garcia (2012) describe this definite position within a group by means of *similarity* relations between the individual and the *rest* of the group. They emphasize that “the key conceptual and psychological contrast in groups is between self and others and not between self and group” (Kenny & Garcia, 2012, p. 471). Indeed, people primarily perceive themselves not as contrary to a group average but rather as opposites to the rest of a group. Consequently, for specific teachers, the homogeneity and heterogeneity of their group always take the form of similarities between themselves and the others in their group into account. Kenny and Garcia (2012) proposed to model such an inclusion of separate positions within a group and their

similarities with the rest of their group using the Group Actor-Partner Interdependence Model (GAPIM), which will be outlined in the following section.

6.4 Modelling Position Effects

Using the GAPIM, the individual value of an interesting feature of a group member is conceived as the result of four different terms or predictors: *actor effect* X , *others' effect* X' , *actor similarity* I , and *others' similarity* I' . A group member is defined as the actor and the rest of the group as the others. The actor effect designates the influence of an independent variable of a group member on its dependent variable, for example the influence of self-efficacy on one's own level of satisfaction. The others' effect then designates the influence of the average of the same independent variable of the others on the dependent variable of the actor. With these two main effects, Kenny, Mannetti, Pierro, Livi, and Kashy (2002) revised the classical multilevel analysis. Group effect, or influence of the group level, is not included as usual in the analysis as total group value; only the average value of the others is included in the GAPIM. In doing so, the influence of the actor is partialized out of the group value.

In addition to the two main effects, *actor effect* and *others' effect*, there are two similarity effects for the study of composition effects. These are based on actor similarity, which models the similarity between the actor and every single other group member regarding an independent variable. Others' similarity models how similar the others are to each other. These similarity terms represent values for the respective position of the actor within the group regarding the independent variable. In addition, these values can now be entered into the analysis as well, whereby the influence of the similarity between actor and others, and among the others, on the dependent variable of the actor can be calculated. In this way, a group composition *from the perspective* of each group member can be modelled. Hence, a value on the individual level is predicted on the basis of two main effects and two similarity effects. If the level of actor similarity is high, the actor is in a numerically more dominant subgroup or in a more homogeneous overall group; if it is low, the actor is isolated from the rest of the group, or at least from every single other in the group. If the level of others' similarity is high, the rest of the group is homogeneous and forms a dominant subgroup, or a homogeneous overall group together with the actor. For an extremely isolated teacher, there is low actor similarity and high others' similarity; thus, the teacher is confronted with a homogeneous, numerically dominant subgroup, of which he or she is not a member. In contrast, when there is high actor similarity and high others' similarity, then the teacher is part of a homogeneous subgroup.

According to Kenny and Garcia (2012), an individual value of a dependent variable (Y_{ik}) consists computationally of a constant (b_{0k}), the four outlined effects ($b_1X_{ik}; b_2X'_{ik}; b_3I_{ik}; b_4I'_{ik}$), and an error term (e_{ik}):

$$Y_{ik} = b_{0k} + b_1X_{ik} + b_2X'_{ik} + b_3I_{ik} + b_4I'_{ik} + e_{ik}$$

Note that $b_2X'_{ik}$, b_3I_{ik} and $b_4I'_{ik}$ constitute effects that relate to the others in the group or to the teacher's relation to the others in the group. Therefore, they are included computationally on the individual level in the present analysis.

In addition, to examine socio-psychological group theories, the four terms can be coded in such a way that different group compositions can be estimated by contrasts, fixations, and equations and compared with each other via model fit (Kenny & Garcia, 2012). With these submodels, it can be determined to which features group members react more sensitively regarding composition effects in general. Accordingly, the two main effects can be analysed in a *Main Effects Model*; the actor effects can be solely analysed in the *Actor Only Model*; and the others' effects can be solely analysed in an *Others Only Model*. In the *Group Model*, actor and others' effects are equated with each other, whereby this model represents the classical multilevel model. Finally, in the *Main Effects Contrast Model*, actor and others' effects are contrasted.

The inclusion of similarity effects thus allows for more differentiated modelling possibilities than have been available up to now. In a *Person-Fit Model*, where the suitability of the separate group member regarding the rest of the group matters, the inclusion of actor similarity in addition to the main effects leads to the best model fit. In a *Diversity Model*, where diversity in the whole group matters, the inclusion of both similarity effects in addition to the main effects leads to the best model fit. In a *Complete Contrast Model*, where the contrast between actor similarity and others' similarity matters, the complementary coding of the similarity effects in addition to the main effects leads to the best model fit. Finally, if all four terms are included without constraints, we refer simply to a *Complete Model*.

6.5 Present Study: The Relation Between the Influence of Composition and Similarity Effects on Job Satisfaction

The advantages of the GAPIM over a conventional multilevel analysis will be illustrated by means of an example from school research. Based on a data set from a study on the effects of the introduction of state-wide exit examinations on schools, teachers, and students (ISCED 3a) (Maag Merki, 2012), we analyse how motivational characteristics of teachers – individual teacher self-efficacy (ITE) and perceived collective teacher self-efficacy (CTE) – affect job satisfaction. With this, we focus on an example that deals with teachers at the individual level and with the teaching staff of the school at the group level. We calculate the influences of the main effect on the group level (group mean), the composition effect on the group level (standard deviation), the main effects on the individual level (actor effect and others' effect), and the position effects on the individual level (actor similarity and others' similarity) on individual job satisfaction.

The two self-efficacy variables qualify for the GAPIM for two reasons: First, in accordance with 'big-fish-little-pond effect' research (Marsh et al., 2008), it can be

assumed that motivational characteristics are especially sensitive to composition and positioning effects because comparison processes with the 'others' are crucial. Second, the two self-efficacy variables share a conceptual similarity, albeit on different levels (individual and group level).

The two concepts, ITE and CTE, refer to Banduras' (1997) concept of self-efficacy. They both describe the individual's perception of being able to master future challenges (Schmitz & Schwarzer, 2002). However, ITE describes the perceived abilities and potentials of the separate teachers, whereas CTE describes the teaching staff's collective self-efficacy, which is perceived and assessed on an individual level as well (Goddard, Hoy, & Hoy, 2000; Schwarzer & Jerusalem, 2002). According to Schwarzer and Jerusalem (2002), CTE consists of meta-individual beliefs of the teaching staff concerning being able to manage future events in a positive manner as a team. ITE and CTE correlate with each other, but they can be described as independent constructs because of their only moderately high level of correlation (Schmitz & Schwarzer, 2002). The question arises here as to what extent CTE really represents meta-individual beliefs or whether it only represents ITE at its own level (Schwarzer & Schmitz, 1999; Skaalvik & Skaalvik, 2007).

According to group main, group composition, and individual main and positioning effects explained above, there are three ways that ITE and CTE can have an effect on job satisfaction.

First, self-efficacy beliefs generally exhibit a positive correlation with job satisfaction. Positive correlations have been found regarding general self-efficacy (Judge & Bono, 2001), individual teacher self-efficacy (ITE) (Caprara, Barbaranelli, Borgogni, & Steca, 2003; Klassen, Usher, & Bong, 2010), and collective teacher self-efficacy (CTE) (Caprara et al., 2003; Klassen et al., 2010; Skaalvik & Skaalvik, 2007). Therefore, we expect to find direct main effects of ITE and CTE – on both the individual and group level – on individual job satisfaction. Teachers with high ITE and teachers, who perceived high CTE, should have higher individual job satisfaction. And teaching staffs where teachers report on average higher ITE and CTE should lead to higher individual job satisfaction of the teachers.

Second, we also expect composition effects of ITE and CTE on individual job satisfaction. Various studies show that the teachers' perceptions of their own coping resources or the coping resources of their team can vary within a team (e.g. Moolenaar, Slegers, & Daly, 2012; Schmitz & Schwarzer, 2002). Further, schools differ in their composition of teachers regarding ITE (Schwarzer & Schmitz, 1999). If some teachers on the teaching staff report low levels of ITE and CTE, while other teachers show high levels, then this variation could lead to high levels of separation. From an interference-oriented perspective, this could have a negative effect on individual job satisfaction. Separation of ITE can indicate an actual lack of collective problem-solving processes in the teaching staff, and it should therefore be congruent with the perception of low CTE. In addition, separation of CTE indicates not only that there is a lack of collective problem-solving processes, but also that teachers experience their same teaching staff differently. In this case, some teachers believe in their collective ability to master future problems, while other teachers do not. The separation of CTE indicates disagreement on the way of looking at a

problem. Therefore, teachers on teaching staffs with high separation of ITE and CTE could have lower job satisfaction than their counterparts on teaching staffs with homogeneous ITE and CTE reports.

Third, in addition to individual main effects, we expect to find positioning effects of ITE and CTE on the individual level on individual job satisfaction. The fact of being isolated on a teaching staff could decrease individual job satisfaction. This is obvious for teachers with low ITE on a teaching staff with others having high ITE. However, in the opposite case, too – for teachers with high ITE on a teaching staff with others having low ITE – isolation can have negative effects on individual job satisfaction. Sharing the same fate of low ITE can lead to similar perspectives and collective support and can help build trust and ties. Being barred from such a collective support can harm individual job satisfaction. The same holds true for CTE. But additionally, CTE refers to an individual's perception of a collective characteristic. Therefore, when a teacher's perception of CTE differs strongly from the others' perceptions, it can be assumed that this teacher does not share all collective processes of the teaching staff. Referring to CTE, isolation can thus indicate objective isolation within the teaching staff and can be detrimental to individual job satisfaction. Therefore, in terms of the GAPIM, the others' similarity of ITE and CTE should have a negative effect on job satisfaction, and the actor's similarity of ITE and CTE should have a positive effect thereon.

6.6 Methods

6.6.1 Sample

The study took place from 2007 to 2011 in the two German states of Bremen and Hesse, which introduced state-wide exit examinations at the end of secondary school (ISCED 3sa). Standardized surveys were conducted in 2007, 2008, 2009, and 2011 (Maag Merki, 2016). In total, 37 secondary schools participated, and surveys were administered to teachers and students. In Bremen, all but one secondary school took part in the surveys (19 schools). In Hesse, the schools were chosen based on crucial context factors (e.g. region, urban–rural, profile of the school). The current study used the teacher data from 2008, which was the first year in which the teachers in both states had to deal with state-wide exit examinations.² A sufficiently large school sample ($N = 37$) and teacher samples (total $N = 1526$, $N_{\text{Bremen}} = 577$, $N_{\text{Hesse}} = 949$) were available to be used for the multilevel analyses. The response rate was sufficient, at 59%. The composition of the sample can be regarded as being representative for both Hesse and Bremen regarding teacher gender and amount (hours) of teaching activity. Young teachers were somewhat over-represented and

²As mentioned above, the analyses of the effects of the implementation of state-wide exit examinations are not the focus of this paper.

teachers older than 50 slightly under-represented. Further descriptive statistics are available in Merki and Oerke (2012).

6.6.2 *Measurement Instruments*

ITE was collected using a scale by Schwarzer, Schmitz, and Daytner (1999) with six items; the scale exhibited a range of 1 to 4 ($\alpha = .74$; $M = 2.84$; $SD = 0.44$). An example item is: “Even if I get disrupted while teaching, I am confident that I can maintain my composure.” The response scale ranged from 1 = not at all true, 2 = barely true, 3 = moderately true, to 4 = exactly true. Since this scale is skewed, it was transformed into an ordinal variable with four categories.

CTE was measured with five items that exhibited a range of 1 to 4 ($\alpha = .76$; $M = 2.54$; $SD = 0.51$) (Halbheer, Kunz, & Maag Merki, 2005; Schwarzer & Jerusalem, 1999). An example item is: “We as teachers are able to deal with ‘difficult’ students because we have the same pedagogical objectives.” The response scale ranged from 1 = not at all true, 2 = barely true, 3 = moderately true, to 4 = exactly true.

Job satisfaction was assessed with six items that exhibited a range of 1 to 4 ($\alpha = .80$; $M = 1.88$; $SD = 0.51$) (Halbheer et al., 2005). The scale entered the study with z-standardization. An example item on the job satisfaction scale is: “I am enjoying my job.” The response scale ranged from 1 = not at all true, 2 = barely true, 3 = moderately true, to 4 = exactly true.

6.6.3 *Analysis Strategies*

The different theoretical and methodological approaches presented above that consider group characteristics in nested data were compared. For this, we first calculated the measure that is usually considered a requirement for a conventional multilevel analysis, the intraclass correlation (ICC). As described above, ICC states how much of the total variability comes from the variability between teaching staffs and from the variability within teaching staffs. Thus, ICC refers to a limited understanding of non-independence of teacher consensus within a teaching staff. A significant ICC size – tested with the Wald-Z – would then indicate that teachers within a teaching staff are over-proportionally similar. However, a non-significant ICC size would indicate a lack of convergence of teachers and would be interpreted as independence of teachers within a teaching staff. In this case, referring to the conventional procedure, the assumption of nested data would be withdrawn, and there would be no necessity for a multilevel analysis.

Second, we calculated a multilevel analysis to examine, if there was a *main group level effect* of the two self-efficacy variables on the teaching staff level to job satisfaction on the individual level. For this purpose, the group means of ITE

($M = 2.840$; $SD = 0.0949$) and CTE ($M = 2.520$; $SD = 0.1640$) on the teaching staff level were calculated as predictors of job satisfaction on the individual level.

In a third step, we examined if there was a *composition effect* of the two self-efficacy variables on the teaching staff level to job satisfaction on the individual level. In this case, we operationalized composition as separation within the teaching staffs and thus as standard deviation. For this purpose, the standard deviations of ITE ($M = 0.434$; $SD = 0.0651$) and CTE ($M = 0.4813$; $SD = 0.0912$) were calculated on the teaching staff level as predictors of job satisfaction on the separate teacher level.

In a fourth step, we examined *main and similarity individual level effects* on the separate teacher level using the GAPIM. For this purpose, we used Kenny and Garcia's macro for SPSS (Kenny & Garcia, 2012). It is based on the linear mixed model in SPSS. The advantage of the macro is that it automatically calculates main and similarity terms and compares the different submodels with each other according to the fit index SABIC (Sample-size Adjusted Bayesian Information Criterion). In addition, we calculated Chi^2 difference tests to estimate whether some differences between the model fit of submodels were significant; Chi^2 difference tests were based on the log-likelihood values. To calculate the similarity terms, continuous and categorical predictors have to be transformed in such a manner that the lowest value is -1 and the highest value 1 .

For samples in the field, however, the problem of multi-collinearity arises. The main effects tend to covary with the similarity effects regarding skewed predictors. For example, if a sample consists of only a few teachers that scored low on individual self-efficacy, it is more likely that these teachers differ from the other members of the teaching staff, i.e., that the similarity term I is smaller. To counter this confound, the skewed continuous predictor ITE is recoded to an ordinal scale. The continuous variable is divided into quartiles; the new ordinal variable thus consists of four categories with equal amount of cases.

To show the benefits of using the GAPIM, the Actor Only Model is reported with only the main actor effect X . It corresponded to a multilevel model with a predictor variable on the individual level. The Main Effects Model followed by adding the main others effect X' , which describes the average predictor effect of the rest of the teaching staff. In this context, the GAPIM differs from the classical multilevel model because the predictor variable was not included in the analysis on the group level (as group average) but entered the analysis with X' as a variable on the individual level. With the Complete Model, finally, the two similarity terms actor similarity I and others' similarity I' were added, which constitute the specific nature of GAPIM.

6.7 Results

6.7.1 Analysis of Variance

In a first step, we analysed to what extent a multilevel model that follows common criteria is necessary at all regarding the dependent variable job satisfaction. A fully unconditional, or no predictors, model resulted in an insignificant group level variability of 0.01243 with a Wald-Z of 1.540 ($p = .124$) and an intraclass correlation of $ICC = 0.01243$. According to Heck, Thomas, and Tabata (2010), the percentage of variability of the dependent variable that is attributed to the group level is too small to be acknowledged with an ICC value below 0.05.

According to common criteria, a multilevel analysis would be refrained from because it is to be assumed that only a small part of the total variability of job satisfaction is to be attributed to differences between the teaching staffs. As has been argued, this point of view reduces non-independence in nested data to homogeneity within a unit and ignores that non-independence can also be described by specific compositions within units. Refraining from carrying out a multilevel analysis, at this point, could lead to missing information about composition and positioning effects.

6.7.2 Main and Composition Effects

In a second and third step, we analysed the main and composition effects on the teaching staff level on individual job satisfaction. In the linear mixed regression model with group mean of ITE (main effect) and standard deviation of ITE (composition effect) as group level predictors, job satisfaction was predicted only by the group mean, with $B = 0.755$ ($p = .000$). The standard deviation of ITE had no significant effect on job satisfaction ($B = -0.026$; $p = .957$).

The result for CTE was the same: Job satisfaction was predicted by the group mean of CTE (main effect) ($B = 1.151$; $p = .000$). The standard deviation of CTE (composition effect) had no significant effect on job satisfaction ($B = -0.197$; $p = .725$).

Consequently, there are only main and but no composition effects in classical multilevel analyses with predictors on the group level. Teaching staffs with high ITE and CTE levels on average, indeed, showed higher levels of individual job satisfaction. The level of separation between the teachers regarding these variables, however, had no influence on individual job satisfaction.

6.7.3 Main and Similarity Effects with GAPIM and Multilevel Analysis

In a fourth step, we analysed main effects and similarity effects on the individual level on individual job satisfaction.

6.7.3.1 Individual Teacher Self-Efficacy as Predictor

Table 6.1 lists all submodels – the Actor Only Model, the Main Effects Model, and the Complete Model. The Actor Only Model showed that individual job satisfaction was predicted by ITE with $B = .714$ ($p = .000$), and it had a multiple correlation of R^2 of .528. For the Main Effects Model, we included the X' term, i.e. the average ITE of the rest of the teaching staff. But X' had no significant effect, with $B = 0.18$ ($p = .888$). For the Complete Model, we finally included the similarity terms I , i.e. the similarity of the actor compared to the other members of the teaching staff, and I' , i.e. the similarity of the other members of the teaching staff among themselves regarding ITE. The Complete Model showed that ITE still had a positive main effect on the individual level of job satisfaction, with $B = .697$ ($p = .000$). The X' term remained insignificant, with $B = .078$ ($p = .616$), and the I term was insignificant as well, with $B = .210$ ($p = .276$). The I' term had a marginally significant effect, with $B = -1.521$ ($p = .056$), however. This means a teacher’s job satisfaction was the lower, the more the other teachers agreed in their ITE reports. Whenever the other teachers were divided in their ITE reports, then the teacher’s job satisfaction increased. This can be quantified in an example of a teacher on a teaching staff with eleven other teachers: A teacher reported a lower job satisfaction of 1.651 standard deviations while all other teachers reported the same ITE as opposed to when six other teachers reported the lowest ITE and five teachers the highest.

With a lower SABIC of 3656.934 ($R^2 = .529$), the model fit of the Complete Model indeed exceeded the model fit of the Actor Only Model (SABIC = 3660.328,

Table 6.1 Effect coefficient estimations and model fits of ITE on job satisfaction

Model	Main effects		Similarity effects		Model fit	
	X	X'	I	I'	SABIC ^b	R ²
Empty	– ^a	– ^a	– ^a	– ^a	4134.71	.000
Actor only	0.714***	– ^a	– ^a	– ^a	3660.33	.528
Main effects	0.713***	0.018	– ^a	– ^a	3660.79	.528
Complete	0.697***	0.072	0.210	–1.513+	3656.93	.529

Note. X = Actors individual teacher self-efficacy; X' = Others' individual teacher self-efficacy; I = Actor similarity; I' = Others' similarity; SABIC = Sample-size adjusted Bayesian information criterion

+ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

^aFixed to zero

^bSmaller SABIC means a better fitting model

$R^2 = .528$). But the improvement in the model fit was not significant ($Chi^2 = 4.851$; $df = 3$; $p = .183$). However, our primary interest was not in the best fitting model, but in showing that by using the GAPIM, we are able to obtain additional information about positioning effects. In this case, we found that a teacher's job satisfaction was not only positively influenced by its ITE, but was also (in tendency) negatively influenced by the similarity of the rest of the teachers on staff regarding their ITE.

6.7.3.2 Collective Teacher Self-Efficacy as Predictor

Table 6.2 also lists all submodels – the Actor Only Model, the Main Effects Model, and the Complete Model. The Actor Only Model showed that the individual level of job satisfaction was predicted by CTE with $B = 1.356$ ($p = .000$) and had a multiple correlation of R^2 of $.457$. In the Main Effects Model, the additional X' term had no significant effect, with $B = -.180$ ($p = .536$). The Complete Model, finally, showed that CTE still had a positive main effect on the individual level of job satisfaction, with $B = 1.322$ ($p = .000$). The X' term remained insignificant, with $B = 0.115$ ($p = .776$). The I term, i.e. the similarity of the actor to the other members of the teaching staff, was significant, with $B = 1.627$ ($p = .031$), and the I' term was insignificant, with $B = -3.919$ ($p = .128$). This means that a teacher's job satisfaction was the higher, the more similar his or her CTE was to that of the other teachers. This can be quantified: A teacher reported a higher job satisfaction of 3.255 standard deviations, if he or she reported exactly the same CTE as the other teachers on staff than if he or she reported the most divergent CTE compared to other teachers on staff.

With a lower SABIC of 3752.214 ($R^2 = .459$), the model fit of the Complete Model indeed exceeded the model fit of the Actor Only Model (SABIC = 3757.594, $R^2 = .457$), although the improvement in the model fit was only nearly significant ($Chi^2 = 6.837$; $df = 3$; $p = .077$). However, this does not lower the importance of the result that teachers' job satisfaction was positively influenced not only by its CTE, but also by the fact how similar he or she perceived CTE compared to the other teachers on staff.

Table 6.2 Effect coefficient estimations and model fits of CTE on job satisfaction

Model	Main effects		Similarity effects		Model fit	
	X	X'	I	I'	SABIC ^a	R ²
Empty	– ^b	– ^b	– ^b	– ^b	4094.59	.000
Actor only	1.356***	– ^b	– ^b	– ^b	3757.59	.457
Main effects	1.362***	–0.180	– ^b	– ^b	3757.68	.457
Complete	1.332***	0.115	1.627*	–3.919	3752.21	.459

Note. X = Actors individual teacher self-efficacy; X' = Others' individual teacher self-efficacy; I = Actor similarity; I' = Others' similarity; SABIC = Sample-size adjusted Bayesian information criterion

+ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

^aSmaller SABIC means a better fitting model

^bFixed to zero

6.8 Discussion

In this contribution, we have argued that especially in the field of school improvement research, composition effects should be taken into consideration for the analysis of nested data. And, thus, in multilevel analysis of nested data in school research, it is necessary that the double character of school levels or classroom levels be disentangled as a result of both the global property of a group level – a separate area of responsibility or shared context – and the collective group composition. Furthermore, non-independence and shared higher-level context in nested data do not necessarily result in similar and converging lower level reports – namely, in shared properties – but can also result in a specific configural group property. Therefore, we discussed advances in research on small groups and organizations to present a differentiated model of the double character of group levels in the school environment. We then discussed different types of diversity (separation, variety, and disparity) to describe the composition of a group (in this case, the teaching staff). Methodically, this leads to the necessity of multilevel analyses to include, apart from group means, statistical diversity measures as predictors, such as standard deviation. We then argued that these composition effects could be translated into positioning effects for the individuals of a group because each individual takes a specific position in the composition of a group. The specific individual position can only be described while accounting for the others in the group and in relation to those others. This leads to the methodological proposition of the GAPIM, which provides additional effect terms to conventional multilevel analyses. The others in the group are accounted for with their average values and their similarity among each other as predictors. Further, the relation to those others is accounted for with the similarity of the actor to the others as a predictor. Therefore, the GAPIM allows for the calculation of the effects of the position of individuals within a group regarding an independent variable on an individual dependent variable. We demonstrated the methodological implementation of the GAPIM exemplarily by analysing individual and collective teacher self-efficacy effects on teachers' individual job satisfaction.

The application of the GAPIM has clear advantages over classical multilevel analyses. To begin with, the necessity of multilevel models is usually determined by the presence of a high ICC. The ICC estimates what part of the total variability of a dependent variable is explained by differences between groups and is thus a measurement of the converging influence that a group has on its members. Therefore, with a lower ICC, there would be no assumed nested structure of the data set, and therefore, no further multilevel analysis would be carried out. In our example, a lower ICC was reported regarding job satisfaction, after which further consideration of teaching staff or the group levels would have been obsolete. Including the GAPIM, however, revealed positioning effects that could not be uncovered without considering the nested structure of the data.

The inclusion of the standard deviation as a group composition measure in a multilevel analysis showed no effects of ITE or CTE. In this case, separation of self-efficacy within a group seems to have no effect on the individual level of job

satisfaction. In other words, a teacher's individual job satisfaction does not seem to depend on whether he or she is in a homogeneous or in a highly split teaching staff regarding individual and collective teacher self-efficacy. From a theoretical point of view, it would not have been sensible to conceptualize the diversity of ICE and CTE as variety or disparity. As for other variables in multilevel analyses, Blau's index for variety, or the proportional relation between group members and resources for disparity, could have been included in the same manner as the standard deviation has been. Therefore, this method is promising for formulating questions on different diversity types and providing additional information about composition effects.

Subsequently, the results of the GAPIM showed that position effects of ITE and CTE, indeed, had effects on teachers' individual job satisfaction. In the GAPIM, group composition was translated into position effects by using similarity measures. Similarity measures describe how strongly the actor corresponds with the others in the group regarding the independent variable, as the term I , or how much the rest of the group resembles itself regarding the independent variable, as the term I' .

Regarding ITE, we found that a teacher's job satisfaction was higher, the higher his or her ITE was (main effect of X). However, there is a tendency that job satisfaction was lower, the more the other teachers on staff related to each other regarding their individual self-efficacy (similarity effect of I'), i.e. the homogeneity of the other teachers on staff lowered the measure of influence of individual self-efficacy on job satisfaction (in tendency). *Nota bene*: This effect remained independent, regardless of whether or not the other teachers on staff reported homogeneously high or homogeneously low ITE; it also remained independent, regardless of whether the actor, i.e. a separate teacher, was a part of this homogeneity or not. Since there was no similarity effect I to be found, we have come to know that the similarity of the actor to the other teachers on staff was not important for individual job satisfaction. For individual job satisfaction to occur, it is preferable for a teacher to work together with other teachers who are diverse in their ITE. This becomes transparent, if you consider that, if there is too high homogeneity regarding the individual estimation of ITE, this can limit the possibilities to enter into an exchange with other teachers concerning individual self-efficacy. Individual job satisfaction may decrease, if the rest of a group perceives and acts monolithically.

Regarding CTE, we found that a teacher's individual job satisfaction was higher, the higher collective self-efficacy was as reported by the teacher (main effect of X). In addition, job satisfaction was higher, the more similar the teacher's estimation regarding collective self-efficacy was to the estimation by the rest of the group (similarity effect of I). *Nota bene*: This effect remained independent, regardless of whether or not a teacher's estimated CTE was similarly high or low to his or her colleagues' estimates. Furthermore, the results showed that it was not the average value of the estimations of CTE by the other teachers on staff that had an influence on individual job satisfaction. Therefore, the fact alone that a teacher exhibits a similar estimation as his or her fellow teachers on staff, increases his or her job satisfaction. This can be interpreted as an integration effect. Regardless of how high the estimations are that refer to the shared estimation of CTE, the integration of a shared estimation affects job satisfaction in a positive manner. In contrast, teachers,

who are isolated because of their CTE estimations, show rather low job satisfaction.

Both examples offer arguments supporting the fact that it is not only one's individual and collective teacher efficacy that is of importance for job satisfaction, but also the similarity that prevails within a teaching staff. Yet, the examples imply as well that these similarity effects exhibit complex dynamics. In the case of individual self-efficacy, the similarity of the other teachers on staff decreases a teacher's job satisfaction. This may be explained from a resource-oriented perspective on diversity. Working in a teaching staff, where the other teachers express diverse levels of individual self-efficacy, makes it apparent that individual self-efficacy is alterable and can be affected by different teaching experiences. This could motivate the separate teacher to question work routines and habits and to improve teaching and professionalisation and, thus, lead to higher job satisfaction. In contrast, when the other teachers express a homogeneous level of individual self-efficacy, a teacher could underestimate the possibility of changing work routines and habits and accept his or her individual self-efficacy level as unalterable. Therefore, diversity in individual self-efficacy would be a resource because it serves as a cue to alterable and diverse experiences. In the case of separately perceived collective self-efficacy, the similarity of a teacher to the rest of the teaching staff increases a teacher's job satisfaction. This may be explained from an interference-oriented perspective on diversity. Collective teacher efficacy is meant to be a shared phenomenon and, thus, should be perceived on a similar level by the teachers involved. Therefore, deviations of a separate teacher's perception from the other teachers' perceptions indicate interferences in the group process. Disagreement on a shared foundation can lead to lower job satisfaction.

Therefore, although composition effects on the teaching staff level could not be found, including the GAPIM, research revealed that the composition of a group has an effect on individual job satisfaction through the position of the individual and the individual's similarity relations to the rest of the group. Introducing the GAPIM into school improvement research, then, can provide additional information. Self-evidently, this fact also applies to other unit levels, such as the classroom. Using this method, loneliness and popularity (Gommans et al., 2017; Gommans, Lodder, & Cillessen, 2016) and academic self-concept (Zurbriggen, Gommans, & Venetz, 2016) have been analysed at the classroom level.

6.9 Limitations and Further Research

Despite the theoretically deduced necessity to take composition effects into account, and despite the empirical results that showed that differences between individuals can be explained in a better way by considering additional information on an individual and group level, there are certain difficulties to be expected regarding the implementation of the GAPIM in the field of school improvement research. In field research, we are interested in independent variables that likely have a skewed

distribution. Thereby, it is to be assumed that the multi-collinearity of the different GAPIM terms presents a problem, and this limits the applicability of similarity effects for the analysis. In this contribution, we managed to avoid collinearity by transforming the continuing variables into categorical variables. In addition, the analyses realized in this contribution are limited to cross-sectional data. It would be interesting, for example, to analyse to what extent composition and similarity have an effect on the changes of separate features, e.g. job satisfaction. Further studies need to be conducted in order to examine to what extent dimensions regarding school efficiency and school development are sensitive to composition and similarity effects. Additionally, complementary analyses, such as social network analyses, could increase the benefits of the presented analyses. These analyses are able to make the collective structures and dynamics visible, for example a collective's density or reciprocal relations, and to develop information for the GAPIM regarding the individuals within the collective, for example a person's in- and out-centrality.

In school improvement research, it is widely acknowledged that the school environment has a nested data structure and that diversity within units – in particular within a teaching staff – is of interest. However, this acknowledgment usually does not lead to a differentiated description of how units and groups are composed, what effects such compositions can have, and how such composition effects can be accounted for in statistical methods. In this article, we presented theoretical considerations on the double character of group levels and on the conceptualization of group composition and diversity. In this context, we proposed the methodological advancement of the GAPIM to address this important lack in school improvement research. The example application of the GAPIM to composition and positional effects of individual and collective teacher self-efficacy on job satisfaction showed how the GAPIM can be used in school improvement research and what additional information can be expected.

References

- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.
- Blau, P. M. (1977). *Inequality and heterogeneity: A primitive theory of social structure*. New York, NY: Free Press.
- Caprara, G. V., Barbaranelli, C., Borgogni, L., & Steca, P. (2003). Efficacy beliefs as determinants of teachers' job satisfaction. *Journal of Educational Psychology, 95*, 821–832. <https://doi.org/10.1037/0022-0663.95.4.821>
- Fend, H. (2005). *Neue Theorie der Schule*. Wiesbaden, Germany: Springer.
- Fend, H. (2008). *Schule gestalten: Systemsteuerung, Schulentwicklung und Unterrichtsqualität* (1. Aufl.). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften/GWV Fachverlage, Wiesbaden.
- Goddard, R. D., Hoy, W. K., & Hoy, A. W. (2000). Collective teacher efficacy: Its meaning, measure, and impact on student achievement. *American Educational Research Journal, 37*, 479–507.

- Gommans, R., Lodder, G., & Cillessen, A. H. N. (2016). *Effects of classroom likeability composition on adolescent loneliness: A brief introduction to the Group Actor-Partner Interdependence Model (GAPIM)(SRA)*. Utrecht, The Netherlands: Utrecht University Repository.
- Gommans, R., Müller, C. M., Stevens, G. W. J. M., Cillessen, A. H. N., Bogt, T., & Tom, F. M. (2017). Individual popularity, peer group popularity composition and adolescents' alcohol consumption. *Journal of Youth and Adolescence*, *46*, 1716–1726. <https://doi.org/10.1007/s10964-016-0611-2>
- Halbheer, U., Kunz, A., & Maag Merki, K. (2005). *Pädagogische Entwicklungsbilanzen an Zürcher Mittelschulen. Indikatoren zu Kontextmerkmalen gymnasialer Bildung. Perspektive der Lehrpersonen: Schul- und Unterrichtserfahrungen. Skalen- und Iemdokumentation*. Zürich, Switzerland, Forschungsbereich Schulqualität & Schulentwicklung, Pädagogisches Institut, Universität Zürich.
- Hallinger, P., & Heck, R. H. (1998). Exploring the principal's contribution to school effectiveness: 1980–1995. *School Effectiveness and School Improvement*, *9*, 157–191. <https://doi.org/10.1080/0924345980090203>
- Hargreaves, A., & Shirley, D. (2009). *The fourth way: The inspiring future for educational change*. Thousand Oaks, CA: Corwin, a Sage Company.
- Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *The Academy of Management Review*, *32*(4), 1199–1228. <https://doi.org/10.2307/20159363>
- Heck, R. H., Thomas, S. L., & Tabata, L. N. (2010). *Multilevel and longitudinal modeling with IBM SPSS*. New York, NY: Routledge.
- Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict, and performance in workgroups. *Administrative Science Quarterly*, *44*, 741–763.
- Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits – self-esteem, generalized self-efficacy, locus of control, and emotional stability – with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology*, *86*, 80–92. <https://doi.org/10.1037/0021-9010.86.1.80>
- Kenny, D. A., & Garcia, R. L. (2012). Using the actor–partner interdependence model to study the effects of group composition. *Small Group Research*, *43*, 468–496.
- Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, *83*, 126–137. <https://doi.org/10.1037/0022-3514.83.1.126>
- Klassen, R. M., Usher, E. L., & Bong, M. (2010). Teachers' collective efficacy, job satisfaction, and job stress in cross-cultural context. *Journal of Experimental Education*, *78*, 464–486. <https://doi.org/10.1080/00220970903292975>
- Kozlowski, S. W. J. (2012). Groups and teams in organizations: Studying the multilevel dynamics of emergence. *Methods for Studying Small Groups: A Behind-the-Scenes Guide*, 260–283.
- Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.
- Lau, D. C., & Murnighan, J. K. (1998). Demographic diversity and faultlines: The compositional dynamics of organizational groups. *The Academy of Management Review*, *23*, 325–340.
- LeBreton, J. M., & Senter, J. L. (2007). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852. <https://doi.org/10.1177/1094428106296642>
- Louis, K. S., Marks, H. M., & Kruse, S. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal*, *33*, 757–798. <https://doi.org/10.3102/00028312033004757>
- Luyten, J. W., & Sammons, P. (2010). Multilevel modelling. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 246–276). New York, NY: Routledge. Retrieved from [https://research.utwente.nl/en/publications/multilevel-modelling\(aff00483-96af-4c5a-95a2-831655655d68\).html](https://research.utwente.nl/en/publications/multilevel-modelling(aff00483-96af-4c5a-95a2-831655655d68).html)

- Maag Merki, K. (Ed.). (2012). *Zentralabitur: Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland*. Wiesbaden, Germany: Springer.
- Maag Merki, K. (2016). Die Einführung zentraler Abiturprüfungen in Bremen und Hessen. In J. Kramer, M. Neumann, & U. Trautwein (Eds.), *Abitur und Matura im Wandel: historische Entwicklungslinien, aktuelle Reformen und ihre Effekte* (pp. 129–159). Wiesbaden, Germany: Springer.
- Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish–little-pond effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review*, 20, 319–350. <https://doi.org/10.1007/s10648-008-9075-6>
- Mathieu, J., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997–2007: A review of recent advancements and a glimpse into the future. *Journal of Management*, 34, 410–476. <https://doi.org/10.1177/0149206308316061>
- Merki, K. M., & Oerke, B. (2012). Methodische Grundlagen der Studie. In K. Maag Merki (Ed.), *Zentralabitur: Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (pp. 45–61). Wiesbaden, Germany: Springer.
- Mitchell, C., & Sackney, L. (2000). *Profound improvement: Building capacity for a learning community*. Lisse, The Netherlands: Swets & Zeitlinger B.V.
- Moolenaar, N. M., Slegers, P. J. C., & Daly, A. J. (2012). Teaming up: Linking collaboration networks, collective efficacy, and student achievement. *Teaching and Teacher Education*, 28, 251–262.
- Schmitz, G. S., & Schwarzer, R. (2002). Individuelle und kollektive Selbstwirksamkeitserwartung von Lehrern. *Zeitschrift Für Pädagogik Beiheft*, 44, 192–214.
- Schudel, K. (2012). *Ein Emergenz-basierter Ansatz für die Diversitätsforschung: Soziale Kategorisierung als Reflexion von Verhaltensdiversität*. Unpublished manuscript. https://www.researchgate.net/publication/330116922_Ein_Emergenz-basierter_Ansatz_fur_die_Diversitätsforschung_-_Soziale_Kategorisierung_als_Reflexion_von_Verhaltensdiversitat. <https://doi.org/10.13140/RG.2.2.11603.63529>
- Schwarzer, R., & Jerusalem, M. (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Berlin, Germany: <http://www.psyc.de/skalen-doku.pdf>. Accessed 21 Jan 2019.
- Schwarzer, R., & Jerusalem, M. (2002). Das Konzept der Selbstwirksamkeit. *Zeitschrift Für Pädagogik Beiheft*, 44, 28–53.
- Schwarzer, R., & Schmitz, G. S. (1999). Skala zur Lehrer-Selbstwirksamkeitserwartung. *Skalen Zur Erfassung Von Lehrer- und Schülermerkmalen*, 60–61.
- Schwarzer, R., Schmitz, G. S., & Daytner, G. T. (1999). *The teacher self-efficacy scale*. http://userpage.fu-berlin.de/~health/teacher_se.htm. Accessed 7 Jan 2019.
- Skaalvik, E. M., & Skaalvik, S. (2007). Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. *Journal of Educational Psychology*, 99, 611–625. <https://doi.org/10.1037/0022-0663.99.3.611>
- Stoll, L. (2009). Capacity building for school improvement or creating capacity for learning? A changing landscape. *Journal of Educational Change*, 10, 115–127. <https://doi.org/10.1007/s10833-009-9104-3>
- Van Knippenberg, D., de Dreu, C. K. W., & Homan, A. C. (2004). Work group diversity and group performance: An integrative model and research agenda. *Journal of Applied Psychology*, 89, 1008–1022. <https://doi.org/10.1037/0021-9010.89.6.1008>
- Van Knippenberg, D., & Schippers, M. C. (2006). Work group diversity. *Annual Review of Psychology*, 58, 515–541. <https://doi.org/10.1146/annurev.psych.58.110405.085546>
- Zurbriggen, C., Gommans, R., & Venetz, M. (2016). *The big-fish-little-pond effect on academic self-concept: A comparison of GAPIM and a latent-manifest contextual model (SRA)*. Utrecht, The Netherlands: Utrecht University Repository.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 7

Reframing Educational Leadership Research in the Twenty-First Century



David NG

7.1 Introduction

Educational leadership research has come of age. From its fledgling start in 1960s under the overarching research agenda of educational administration for school improvement, the focus shifted to leadership research from the early 1990s (Boyan, 1981; Day et al., 2010; Griffiths, 1959, 1979; Gronn, 2002; MacBeath & Cheng, 2008; Mulford & Silins, 2003; Southworth, 2002; Witziers, Bosker, & Kruger, 2003). Since then, educational leadership as a respected field began to flourish by the early 2000s (Hallinger, 2013; Robinson, Lloyd, & Rowe, 2008; Walker & Dimmock, 2000). From the 1980s up to the present time, the body of knowledge on educational leadership has grown tremendously to produce three distinctive educational leadership theories: instructional leadership, transformational leadership, and distributed leadership. While it is undisputed that educational leadership research has indeed been productive, there is a sense that a narrowing labyrinth of researchable questions is approaching in particular to the first two educational leadership research theories. The evidence of this is implied in the concerted call to expand and situate educational leadership research in non-Western societies (Dimmock, 2000; Dimmock & Walker, 2005; Hallinger, 2011; Hallinger, Walker, & Bajunid, 2005). This call is valid in that there is still limited contribution to substantive theory building from non-Western societies. However, it also implies that Western societies' focus on educational leadership has reached an optimum stage in publications and knowledge building. A more pertinent reason to rethink educational leadership research could be based on epistemological questions about the social science research paradigm that has been the foundation of educational leadership research.

D. NG (✉)

National Institute of Education, Nanyang Technological University, Singapore, Singapore
e-mail: david.ng@nie.edu.sg

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_7

107

These questions will be expanded as the discussion proceeds on current approaches of educational leadership research.

This chapter has three goals: The first one is to map the data-analytical methods used in educational leadership research over the last thirty years (1980–2016). This investigation covers the research methodologies used in instructional leadership, transformational leadership, and distributed leadership.

Educational leadership studies are conducted in the social context of the school. This context involves complex social interactions between and among leaders, staff, parents, communities, partners, and students. In the last decade, there has been a consensus among scholars that schools have evolved to become more complex. Furthermore, there is a consensus among scholars to view complexity through increases in the number of actors and the interactions between them. The complexity of schools is evident in the rise in accountability and involvement from an expanding number of stakeholders involved, such as politicians, clinical professionals (who diagnose learning disabilities of students), communities, and educational resource providers (training and certifying institutions). The relations between stakeholders are non-linear and discontinuous, so even small changes in variables can have a significant impact on the whole system. Therefore, the second goal is to determine whether methodologies that are adequate for the assessment of complex interaction patterns, influences, interdependencies, and behavioural outcomes that are associated with the social context of the school, have been adopted over the past three decades.

The third goal is to explore potential methodologies in the study of educational leadership. These alternative methodologies are taken from more recent developments of research methodologies used in other fields. These fields, such as health, development of society, among others, have similarities with the study of educational leadership. The common link is the social contexts and the system's influence involving the spectrum of interactions, change, and emergence. We will examine published empirical research and associated theories that look at influence, interdependencies, change, and emergence. Adopting these alternative methodologies will enable reframing educational leadership so it can move forward. Three questions guide the presentation of this paper:

- What are the data sources and analytical methods adopted in educational leadership research?
- What is the current landscape of schooling and how does it challenge current educational leadership research methodologies?
- What are some possible alternative research methodologies and how can they complement current methodologies in educational leadership research?

This chapter proposes to reframe educational leadership studies in view of new knowledge and understanding of alternative research data and analytical methods. It is not the intent of the paper to suggest that current research methodologies are no longer valid. On the contrary, the corpus knowledge of current social science research methodologies practiced, taught, and learned through the past three decades cannot be dismissed lightly. Instead of proposing to reframe educational leadership

studies, the main purpose of this paper is to explore and propose complementary research methodologies that will open up greater opportunities for research investigation. These opportunities are linked to the functions of adopting alternate analytical research tools.

7.2 What Are the Dominant Methodologies Adopted in Educational Leadership Research?

Educational leadership research adopts a spectrum of methods that conform to the characteristics of disciplined inquiry. Cronbach and Suppes (1969) defined disciplined inquiry as “conducted and reported in such a way that the argument can be painstakingly examined” (p. 15). What this means is that any data collected and interpreted through reasoning and arguments must be capable of withstanding careful scrutiny by another research member in the field.

This section looks at the disciplined inquiry methods adopted and implemented in the last thirty years that have contributed to the current body of knowledge on educational leadership and management. The pragmatic rationale to impose a time frame for the review is that instructional leadership was conceptualized in the 1980s, followed by transformational leadership and in recent years, distributed leadership. The purpose of this review is to identify, if possible, all quantitative and qualitative methods adopted. The next section provides a broad overview of the three educational leadership theories/models. This will anchor the discussion on alternate research methodologies that will reframe and expand the research on these theories/models.

7.2.1 Instructional, Transformational, and Distributed Leadership

Instructional leadership became popular during the early 1980s. There are two general concepts of instructional leadership – one is narrow while the other is broad (Sheppard, 1996). The narrow concept defines instructional leadership as actions that are directly related to teaching and learning, such as conducting classroom observations. This was the earlier conceptualization of instructional leadership in the 1980s, and it was normally applied within the context of small, poor urban primary schools (Hallinger, 2003; Meyer & Macmillan, 2001). The broad concept of instructional leadership includes all leadership activities that indirectly affect student learning, including school culture, and time-tabling procedures by impacting the quality of curriculum and instruction delivered to students. This conceptualization acknowledges that principals, as instructional leaders, have a positive impact on students’ learning, but that this influence is mediated (Goldring & Greenfield, 2002;

Leithwood & Jantzi, 2000; Southworth, 2002). A comprehensive model of instructional leadership was developed by Hallinger and Murphy (1985, 1986). This dominant model proposes three dimensions of the instructional leadership construct: defining the school's mission, managing the instructional program, and promoting a positive school-learning climate. Hallinger and Heck (1996), in their comprehensive review of research on school leadership, concluded that instructional leadership was the most commonly researched. The authors' focused review found that over 125 empirical studies employed this construct between 1980 and 2000 (Hallinger, 2003). In the last decade, instructional leadership has regained prominence and attention in part because of the lack of empirical studies in non-Western societies. This can also be inferred from the notion that leadership in curriculum and instruction still matters and remains the core business of schools.

Transformational leadership was introduced as a theory in the general leadership literature during the 1970s and 1980s (e.g. Bass, 1997; Howell & Avolio, 1993). Transformational leadership focuses on developing the organisation's capacity and commitment to innovate (Leithwood & Duke, 1999). Correspondingly, transformational leadership is supposed to enable change to occur (Leithwood, Tomlinson, & Genge, 1996). Amongst the leadership models, transformational leadership is the one most explicitly linked to the implementation of change. It quickly gained popularity among educational leadership researchers during the 1990s in part because of reports of underperforming schools as a result of top-down policy driven changes in the 1980s. Sustained interest during the 1990s was also fuelled by the perception that the instructional leadership model is a directive model (Hallinger & Heck, 1996). In a pointed statement of the extent of instructional leadership research, Hallinger (2003, p. 343) emphatically notes that "The days of the lone instructional leader are over. We no longer believe that one administrator can serve as the instructional leader for the entire school without the substantial participation of other educators." From the beginning of the 2000s, a series of review studies comparing the effects of transformational leadership and instructional leadership, the 'over-prescriptivity' of findings, the limited methodologies adopted, and a lack of international research contributed to the waning interest in transformational leadership (Robinson et al., 2008; Robinson, 2010).

Interest in distributed leadership took off at around 2000. Gronn (2002), and Spillane, Halverson, and Diamond (2004) are leading the current debate on distributed leadership as observed by Harris (2005). Gronn's concept of distributed leadership is a "purely theoretical exploration" (p. 258) while Spillane's and his various colleagues' work is based on empirical studies that are still ongoing. When Gronn and Spillane first proposed their concepts of distributed leadership, what was revolutionary was a shift from focusing on the leadership actions of an individual as a sole agent to analyzing the 'concertive' or 'conjoint' actions of multiple individuals interacting and leading within a specific social and cultural context (Bennett, Wise, Woods, & Harvey, 2003; Gronn, 2002, 2009; Spillane, 2005; Woods, 2004). In addition, Spillane, Diamond, and Jita (2003) explicitly relate their concept of distributed leadership to instructional improvement, which, therefore, catalyzes the interest among researchers to explore the constructs in school improvement and

effectiveness. From 2000 to 2016, a focused search for empirical studies that employed the constructs of distributed leadership yielded over 97 studies.

7.2.2 Assessment of the Dominant Methodologies in Educational Leadership Research and Courses

The purpose of this review is to identify, if possible, all the quantitative and qualitative methods adopted. This review is based on a combined search for the three educational leadership theories in schools using the following search parameters:

- Keywords in database search: “instructional leadership” OR “transformational leadership” OR “distributed leadership”
- Limiters: Full Text; Scholarly (Peer-reviewed) Journals; Published Date: 1980–2016
- Narrow by Methodology: quantitative study
- Narrow by Methodology: qualitative study
- Search modes: Find all search terms
- Interface: EBSCOhost Research Databases
- Database: Academic Search Premier; British Education Index; Education Source; ERIC

The search yielded over 672 empirical studies employing the constructs of instructional leadership, transformational leadership, and distributed leadership. As the purpose of the review is to identify all quantitative and qualitative methods adopted, only that information was extracted. The researchers carefully read the relevant sections of the 672 studies pertaining to methodologies and extracted that information. An overview of the results is given in Tables 7.1 and 7.2.

The range of quantitative and qualitative research methodologies and analytical tools found in the review was categorized as follows:

Quantitative Analyses:

- **Univariate Analysis:**
- The analysis refers to a single variable represented by frequency distribution, mean and standard deviation.
- **Bivariate Analysis:**
- This type of analysis examines how two variables are related to each other, represented by ANOVA, Pearson product moment correlations, correlation and regression.
- **Multivariate Analysis:**
- These are statistical procedures that are used to reach conclusions about associations between two or more variables. Representations of inferential statistics include regression coefficients, MANOVA, MANCOVA, two-group comparison (t-test), factor analysis, path analysis, hierarchical linear modeling, and others.

Table 7.1 Quantitative methods used in the study of instructional, transformational, and distributed leadership

Data source:	Types:	Specific analytical methods:
Questionnaire/survey	Basic statistics	Frequency distribution
		Mean
		Median
		Standard deviation
		t-test
	Analysis of variance	Analysis of covariance
		Analysis of variance
		One-way ANOVA
		Two-way ANOVA
	Association and correlation	Correlation
		Regression
	Causal modelling	Dependent variable
		Independent variable
		Path analysis
		Structural equation modelling
	Factor analysis	Exploratory factor analysis
		Factor analysis
		Confirmatory factor analysis
		Oblique rotation
		Rotated factor
	Linear and multilevel analysis	Generalized linear model
		Hierarchical generalized linear model
		Hierarchical linear modelling
	Multilevel regression	Multicollinearity
Multiple regression analysis		
Interaction effect		

Data source: Questionnaire/Survey

Qualitative Analyses:

- Content Analysis:
- Content analysis is the systematic analysis of the text by adopting rules that can separate the text into units of analysis, such as assumptions, effects, enablers and barriers. The text is obtained through document search, artifacts, interviews, field notes, or observations. The transcribed data are converted into protocols followed by categories. Coding schemes are then applied to determine themes and their relations.
- Hermeneutic Analysis:
- With this type of analysis, researchers interpret the subjective meaning of a given text within its socio-historic context. Methods adopted extend beyond texts to encompass all forms of communication, verbal and non-verbal. An iterative analyses method between interpretation of text and holistic under-

Table 7.2 Qualitative methods used in the study of instructional, transformational, and distributed leadership

Data sources:
One-to-one interview
Focus group interview
Document search (e.g. writing samples, e-mail correspondence, and district literature)
Field notes
Classroom observations
Semi-structured interviews
Artifacts
Shadowing
Interview protocols (for multiple case studies)
Interpretive description
Topic-oriented
The voices from the field
Cross-cultural comparative studies
Portfolios
Micro-political analysis
Specific analytical methods:
Thematic analysis (“coding” and then segregating the data by codes into data clumps for further analysis and description)
Discrepancy theme
Characteristics
Descriptive
Factors
Roles
Nature
Content analysis
Causal sequence
Interactions but also in social, cultural, and institutional discourses
Structured coding scheme derived from the conceptual framework
Exploratory analysis
Phenomenology and constant comparative methods
Comparative analysis: Finding common themes, and contrasts
Detailed analytical memo
Vertical analysis: Analyzing participants’ voices separately; and patterns and elucidating the differences among participants’ voices.

standing of the context is adopted in order to develop a fuller understanding of the phenomenon.

- Grounded-theory Analysis:
- This is an inductive technique of interpreting recorded data about a social phenomenon. Data acquired through participant observation, in-depth interviews, focus groups, narratives of audio/video recordings, and documents are interpreted based on empirical data. A systematic coding technique

involving open coding, axial coding, and selective coding is rigorously applied. These coding techniques aim to identify key ideas, categories, and causal relations among categories, finally arriving at a theoretical saturation where additional data and analyses do not yield any marginal change within the core categories.

On the one hand, these results show that a wide range of both quantitative and qualitative methodologies are applied and that the field is open to a lot of diversity in methodologies, but, on the other hand, the results also show that complexity methodology is missing completely.

One of the purposes of this paper is to identify current research methodologies that have been adopted for the past decades. The following review is to ascertain whether current research methodologies adopted are also reinforced and transmitted by the research courses offered by top universities. A search was conducted that specifically looked at graduate research courses taught in educational leadership and management. The following search parameters were used:

- Identify the top 20 universities that offer graduate courses in educational leadership and management.
- QS ranking of universities is chosen over Times ranking because QS ranking is sorted by subject: Education and searchable by educational leadership.
- Representation of Western and Eastern universities in order to provide a representation of universities globally.

The findings are presented in Table 7.3. This table is remarkably similar to Tables 7.1 and 7.2 but with more details of the topics in educational leadership research methodologies. The previously presented findings of the methodologies used in educational leadership research strongly suggest that the research methodologies currently adopted in educational leadership studies are reinforced by research courses taught at the top universities. Indeed, the transmission and application of research skills is a critical and essential component of graduate programmes. This transmission of knowledge and practice is strengthened by the enshrined supervisor-supervisee relationship where cognitive modelling takes place through discourse, reflection, guidance, and inquiry. The one-to-one supervision has the very powerful effect of instilling expectations, cultivating habits, and shaping practices that contribute to a competent researcher identity. It is noteworthy that the transmission-based form has emanated from and is continued in the paradigm of social science. Table 7.3 presents the research courses that are currently taught at the top 20 universities offering educational leadership research.

Table 7.3 Research courses in Educational Leadership taught at the Top 20 universities

Quantitative courses:	Qualitative courses:	Universities:
Basic descriptive measures summarizing data using statistics, such as frequency, mean, and variance;	Content analysis	The UCL Institute of Education
Random sampling and sampling error	Ethnography	Harvard University
Hypothesis tests for continuous and categorical data	Critical ethnography	Stanford University
Modelling continuous data using simple linear regression	Pragmatic qualitative research	University of Cambridge
General linear model: Regression, correlation, analysis of variance, and analysis of covariance	Phenomenological analysis	The University of Melbourne
Multiple linear regression, including categorical covariates and interaction effects, factorial ANOVA, ANCOVA, MANOVA, MANCOVA, partial and semi-partial correlations, path analysis, exploratory factor analysis, and confirmatory factor analysis.	Discourse analysis	The University of Hong Kong
Basic statistical inference, including confidence intervals and hypothesis testing; multiple linear regression, including categorical variables and interaction effects	Analysis of visual materials	University of Oxford
Structural equation modelling	Policy documentary analysis	University of California, LA (UCLA)
SEM with observed variables	Historical documentary analysis	The University of Sydney
SEM with latent variables	Classroom ethnography	Nanyang Technological University
Maximum likelihood estimating, goodness-of-fit measures, nested models	Survey	University of California, Berkeley (UCB)
Binary and multinomial logistic models	Grounded theory	Columbia University

(continued)

Table 7.3 (continued)

Quantitative courses:	Qualitative courses:	Universities:
Instrument reliability and validity	Action research	University of Michigan
	Participatory research	University of Wisconsin-Madison
	Bibliographic analysis	The Hong Kong Institute of Education
	Institutional ethnography	Monash University
	Narrative	University of Toronto
	Observation and interview	University of British Columbia
	Interviews	Michigan State University
	Oral history	The Chinese University of Hong Kong
	Arts-based research	
	Critical transnational ethnography	
	Hermeneutics	
	Phenomenology	
	Semiotics	
Crystallization		

7.3 Limitations of the Dominant Methodologies in Educational Leadership Research and Courses

The range of methodologies and analytical tools reviewed above are disciplined inquiry methods in social science. Social sciences are the science of people or collections of people, such as groups, firms, societies, or economies, and their individual or collective behaviours; social sciences can be classified into different disciplines, such as psychology (the science of human behaviours), sociology (the science of social groups), and economics (the science of firms, markets, and economies). This section is not intended to wade into epistemological and ontological debates within the social sciences. It is also not possible to have an in-depth discussion on social science methodologies within the constraints of this paper. To highlight ongoing discussions about limitations of social science research is the focus of this paper.

Educational leadership is not a discipline by itself, but a field of study that involves events, factors, phenomena, organizations, topics, issues, people, and processes related to leadership in educational settings. This field of study adopts social science inquiry methods. The review of research methodologies, as depicted in

Tables 7.1 and 7.2, strongly suggests that educational leadership research subscribed to the functionalist paradigm (Bhattacharjee, 2012). The functionalist paradigm suggests that social order or patterns can be understood in terms of their functional components. Therefore, the logical steps will involve breaking down a problem into small components and studying one or more components in detail using objectivist techniques, such as surveys and experimental research. It also encompasses an in-depth investigation of the phenomenon in order to uncover themes, categories, and sub-categories.

Educational leadership studies, using quantitative methods, aim to minimize subjectivity. Hence, the constant advocacy of good sampling techniques and a large sample size in order to represent a population where the sample is reported by mean, standard deviation, and normal distribution, among others. Qualitative methods rest upon the assumption that there is no single reality for events, phenomena, and meaning in the social world. Adopting a disciplined analytical method based on dense contextualized data in order to arrive at an acceptable interpretation of complex social phenomena is advocated. The following section will discuss several common limitations of social science research.

7.3.1 Population, Sampling, and Normal Distributions

Based on the review, quantitative and qualitative methods of social science in educational leadership research can be inferred to subscribe to the goals of identifying and analyzing data that can inform about a population. Researchers aim to collect data that either maximize generalization to the population in the case of quantitative methods or provide explanation and interpretation of a phenomenon that represents a population in the case of qualitative methods. In most cases, definitive conclusions of a population are rarely possible in social sciences because data collection of an entire population is seldom achieved.

Therefore, researchers apply sampling procedures where the mean of the sampling distribution will approximate the mean of the true population distribution, which has come to be known as normal distribution. This concept has set the parameters as to how data has been collected and analyzed over many years. It has become widely accepted that most data ought to be near an average value, with a small number of values that are smaller, and the other extreme where values are larger. To calculate these values, the probability density function (PDF), or density of a continuous random variable, is used. It is a function that describes the relative likelihood for this random variable to take on a given value.

A simple example will help to explain this: If 20 school principals were randomly selected and arranged within a room according to their heights, one would most likely see a normal distribution: with a few principals who are the shortest on the left, the majority in the middle, and a few principals who are the tallest on the right. This has come to be known as the normal curve or probability density function.

Most quantitative research involves the use of statistical methods presuming independence among data points and Gaussian “normal” distributions (Andriani & McKelvey, 2007). The Gaussian distribution is characterized by its stable mean and finite variance (Torres-Carrasquillo et al., 2002). Suppose that in the example above the shortest principal is 1.6 m. Given the question, “What is the probability of a principal in the line being shorter than 1.5m?”, the answer would be ‘0’. From the total number of principals in the room, there is no probability to find someone who is shorter than 1.6 m. But if the question were, “What is the probability of a principal in the line being 1.7m?”, then the answer could be 0.2 (i.e. 10%, or 2 persons). Hence, this explains the finite variance, which is dependent upon the sample size. Normal distributions assume few values far from the mean and, therefore, the mean is representative of the population. Even largest deviations, which are exceptionally rare, are still only about a factor of two from the mean in either direction and are well-characterized by quoting a simple standard deviation (Clauset, Shalizi, & Newman, 2009). This property of the normal curve, in particular the notion that extreme ends of variance are less likely to occur, has significant implications as will be discussed.

Is the normal distribution the standard to determine acceptable findings in educational research? One possible answer is a study done by Micceri (1989). His investigation involved obtaining secondary data from 46 different test sources and 89 different populations, and that included psychometric and achievement/ability measures. He managed to obtain analyzed data from 440 researchers; he then submitted these secondary data to analysis and found that they were significantly non-normal at the alpha .01 significance level. In fact, his findings showed that tail weights, exponential-level asymmetry, severe digit preferences, multi-modalities, and modes external to the mean/median interval were evident. His conclusion was that the underlying tenets of normality-assuming statistics appear fallacious for the psychometric measures. Micceri (1989, p. 16) added that “one must conclude that the robustness literature is at best indicative.”

In another well-cited article in the Review of Educational Research, Walberg, Strykowski, Rovai, and Hung (1984, p. 87) state that “considerable evidence shows that positive-skew distributions characterize many objects and fundamental processes in biology, crime, economics, demography, geography, industry, information and library sciences, linguistics, psychology, sociology, and the production and utilization of knowledge.” Perhaps the most pointed statement made by Walberg et al., that “commonly reported univariate statistics such as means, standard deviations, and ranges – as well as bivariate and multivariate statistics [...] and regression weights – are generally useless in revealing skewness” is worthy to note.

What are the implications and limitations of the normal distribution in the population? There are at least two limitations. First, reliance on normal distribution statistics puts a heavy burden on assumptions and procedures. The procedures of randomness and equilibrium have powerful influences on how theories are built and also determine how research questions are formulated. In other words, findings may be rejected that could otherwise be informative because they do not meet the normal distribution litmus. The explanation of the normal distribution suggests that any

events or phenomena at both (extreme) ends of the normal curve are highly unlikely – consequently, we typically reject those findings. Research on real-world phenomena, e.g. social networks, banking networks, and world-wide web networks, has established that events in the tails are more likely to happen than under the assumption of a normal distribution (Mitzenmacher, 2004). Many real-world networks (world-wide web, social networks, professional networks, etc.) have what is known as long-tailed distribution instead of normal distribution.

Second, independent variables contributing to a normal distribution assume that the variables are static. The reality is that in education (and educational leadership) the variables are dynamic. This dynamic function comes from past and even future environmental and individual influences. An example is that of being fortunate to have initial advantages, such as enrolling in a university study (past influence), working with eminent researchers (preferential attachment), obtaining well-funded research projects, and having publication opportunities (environmental influence), combine multiplicatively over time and accumulate to produce a highly skewed number of publications. The distribution would not conform to the normal curve for researchers when past influence, preferential attachment, and environmental influences are taken into consideration. At the moment, the large majority of reviewed studies, using inferential statistics of mean and standard deviations, does not account for such dynamic influences upon the variables. Is there an alternative that could complement this limitation?

7.3.2 Linearity in a Predominantly Closed System

The dominant analytical tools adopted in educational leadership research involve relational and associational analyses of the effects of leadership actions and interventions in schools. The focus is on identifying variables, factors, and their associations in providing explanations of successful practices. The central concept of relations is based on the assumption of linearity. Linearity means two things: Proportionality between cause and effect, and superposition (Nicolis, Prigogine, & Nocolis, 1989). According to this principle, complex problems can be broken down into simpler problems, which can be solved individually. That is, the effects of interventions can be reconstructed by summing up the effects of the single causes acting on the single variable. This, then, allows establishing causality efficiently.

However, this assumption forces researchers to accept that systems are in equilibrium. The first implication is that the number of possible outcomes in a system is limited (because of the limited number of variables within a closed system). The second implication is that moments of instability, such as through an intervention from the school leader, are brief, whereas the duration of the stability of the final outcome is long. In that case, one can measure effects or establish relations, and accept its data value as a true indication of the cause of intervention. For this to be true, however, the many variables in the school (as a closed system) must be assumed to be independent. Other possibilities to this assumption are to have

interdependence, mutual causality, and the occurrence of possible external influences in the larger system (e.g. political or economic change).

The goal of school leadership is to improve student achievement. Student achievement is demonstrable, even though there are considerable differences of opinion about how to define improvement in learning or achievement (Larsen-Freeman, 1997). This is because much research assumes that the classroom is a closed system with defined boundaries, variables, and predictable outcomes. This mechanistic linear view neglects students as active constructors of meaning with diverse views, needs, and goals (Doll Jr, 1989). It is debatable to draw the association directly that teachers' pedagogy results in learning. Luo, Hogan, Yeung, Sheng, and Aye (2014) found that Singapore students attributed their academic success mainly to internal regulations (effort, interest, and study skills), followed by teachers' help, teachers' ability, parents' help, and tuition classes. While the study appears to support linearity and attribute students' academic success to identified variables, there is still much less certainty about other aspects, such as the interaction effects among the variables. The use of generalized linearity cannot account for the interactions among students – how they motivate each other, how they compete, and how they derive the drive to perform. Researchers studying student achievement tend to seek to reduce and consolidate variables in order to discover order while denying irregularity.

Due to its simplicity, linearity became almost universally adopted as the true assumption along with its corresponding measures in educational leadership research. School improvement, student learning, staff capacity, and efficacy are much more complex than directly assigned proportionality between factors and outcomes, and identifying superposition. Cziko (1989, p. 17) asserted that “complex human behaviour of the type that interests educational researchers is by its nature unpredictable if not indeterminate, a view that raises serious questions about the validity of quantitative, experimental, positivist approach to educational research.” In general, school improvement ought to include a notion of and methodology for describing non-linear cognitive systems or processes and to accept that research questions cannot be simplified to find answers from regression models alone, particularly research questions that involve non-specified outcome variables. For instance, school success, in addition to internal variables and factors, simultaneously includes influence by changes in government policies and conflicting demands of multiple stakeholders (e.g. economic and society-related stakeholders). Relying only on the linearity within a closed system will limit any understanding of such interdependencies and mutual influences. Therefore, a holistic and more complete understanding of social phenomena, such as why some school systems in some countries are more successful than others, requires an appreciation and application of research methods that include the elements of open and closed systems. The alternative to linearity – non-linearity, emergence, and self-organization – as an alternate view of reality shall be discussed in the fourth part of this chapter.

7.3.3 *Explanatory, Explorative, and Descriptive Research*

One of the research aims in social science is the understanding of subjectively meaningful experiences. The school of thought that stresses the importance of interpretation and observation in understanding the social situation in schools is also known as ‘interpretivism.’ This is an integral part of qualitative research methodologies and analytical tools adopted in educational leadership research. The interrelatedness of different aspects of staff members’ work (teaching, professional development), interactions with students (learning, guidance, etc.), cultural factors, and others, form a very important focus of qualitative research. Qualitative research practice has reflected this in the use of explanatory, explorative, and descriptive methods, which attempt to provide a holistic understanding of research participants’ views and actions in the context of their lives overall.

Ritchie, Lewis, Nicholls, and Ormston (2013) provide clear explanations for the following research practices: Exploratory research is undertaken to explore an issue or a topic. It is particularly useful in helping to identify a problem, clarify the nature of a problem or define the issues involved. It can be used to develop propositions and hypotheses for further research, to look for new insights or to reach a greater understanding of an issue. For example, one might conduct exploratory research to understand how staff members react to new curriculum plans or ideas for developing holistic achievement, or what teachers mean when they talk about ‘constructivism,’ or to help define what is meant by the term ‘white space.’

A significant number of qualitative studies reviewed in this paper are about description as well as exploration – finding the answers to the Who? What? Where? When? How? and How many? questions. While exploratory research can provide description, the purpose of descriptive research is to answer more clearly defined research questions. Descriptive research aims to provide a perspective for social phenomena or sets of experiences.

Explanatory research addresses the Why questions: Why do staff members value empowerment? Why do some staff members perceive the school climate negatively and others do not? Why do some students have a high self-motivation and others do not? What might explain this? Explanatory, in particular qualitative research assists in answering these types of questions, which allows ruling out rival explanations, guidance to come to valid conclusions, and developing causal explanations.

An obvious limitation of explanatory, explorative, and descriptive educational leadership research is that this is done after an intervention; another limitation constitutes the mere focus on outcomes. If research tapped into this process before interventions were implemented, then two reasonable questions would be:

- Will an intended school vision or policy have the desired positive reception among staff members?
- How can one predict the kind of reception or perception staff members might have?

The answers would be useful for school leaders in order to initiate intervention measures before serious damage occurs. It would be most useful to be able to extrapolate those answers to the larger system, where policy makers are interested in predicting likely outcomes of the policy prior to its implementation. An example of this kind of research is the development of models known as simulations. Computer simulation is known as the third disciplined scientific methodology. This concept will be discussed in the latter section on alternative methodologies.

A summary of the limitations of current methodologies in educational leadership is concisely captured by Leithwood and Jantzi (1999, p. 471): “Finally, even the most sophisticated quantitative designs used in current leadership effects research treat leadership as exogenous variable influencing students, sometimes directly, but mostly indirectly, through school conditions, moderated by student background characteristics. The goal of such research usually is to validate a specific form of leadership by demonstrating significant effects on the school organization and on students. The logic of such designs assumes that influence flows in one direction – from the leader to the student, however tortuous the path might be. But the present study hints at a far more complex set of interactions between leadership, school conditions, and family educational culture in the production of student outcomes.”

7.4 The Current Landscape of Schooling

7.4.1 *Complexity of Schools: Systems and Structures*

Murphy (2015) examined the evolution of education from the industrial era in the USA (1890–1920) to the post-industrial era of the 1980s. He concluded that post-industrial school organizations have fundamentally shifted in roles, relationships, and responsibilities. The shift is seen in the blurring of distinctions between administrators and teachers; general (expanded) roles instead of specialization, where specialization is no longer held in high regard, as compared to the industrial era, with greater flexibility and adaptability. In terms of structures, the traditional hierarchical organizational structures are giving way to structures that are flatter.

This shift in roles, relationships, and responsibilities has (also) contributed to the increasing complexity of schools. The direct and indirect involvement between and among a growing circle of stakeholders within the school and between government, employers, and communities clearly support the view that schooling is no longer seen as a closed system. It is both a closed and open system (Darling-Hammond, 2010; Hargreaves & Shirley, 2009; Leithwood & Day, 2007). Leithwood and Day (2007) state that “Schools are dynamic organizations, and change in ways that cannot be predicted,” as they reviewed leadership studies from eight different countries. Open systems are “a system in exchange of a matter with its environment” (Von Bertalanffy, 1968, p. 141). Schools as an open system are therefore seen as part of a much larger network rather than an independent, self-standing entity.

Thus, to understand the processes still existing within the schools, it is critical to study the interrelations between those entities and their connections to the whole system. The interrelationships among stakeholders are non-linear and discontinuous, so even small changes in variables can have a significant impact on the whole system. This notion of small change leading to global change is reflected in the example of the current ‘world-class education system’ movement. From countries as diverse as the United Arab Emirates, Brazil, Hong Kong, Singapore, Vietnam, Australia, and the United States of America, a common theme found in education reform documents is the term “world-class education.” This term has become widely associated with comparative results on international tests, such as Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA), which purports to measure certain aspects of educational quality. Indeed, the term is frequently used by countries that have attained high scores in these international tests as a strong indicator of being world-class. This seemingly small aspect of change (i.e. the comparing of achievements in Mathematics and Science) has impacted developing and developed nations in reforming their education system and in calling their ongoing education reforms as moving towards a ‘world-class education system.’

Thus, interrelationships in an open system require sophisticated analyses of their systemic nature. A reductionist and linear sequential relationship investigation would not be sufficient in order to bring about further change. To remain of value with the current trends, educational leadership researchers, who adopt complexity methodology, would help practitioners shaping the future by creating an environment of valid knowledge.

7.4.2 Shared and Distributed Leadership

The idea of distributed leadership connects well with the trend towards greater decentralization (since the 1980s) and school autonomy through which school leaders are expected to play a greater role in leadership beyond the school borders and requires them to make budgetary decisions, foster professional capacity development, and play a role in the design of school buildings, and many more aspects (Glatter & Kydd, 2003; Lee, Hallinger, & Walker, 2012; Nguyen, Ng, & Yap, 2017; Spillane, Halverson, & Diamond, 2001).

A core function of leadership – distributed leadership included – is decision-making. The most popular discussion of decision-making of the twenty-first century emanates from the concept of decentralization. Decentralization includes delegating responsibilities, practice of distributed leadership, and practice of distributed or shared instructional leadership (Lee et al., 2012; Nguyen et al., 2017; Spillane et al., 2001).

Glatter and Kydd (2003) identified two models of decentralization, which have important implications for school leaders, namely local empowerment and school empowerment. In local empowerment, the transfer of responsibilities takes place

from the state to the districts, including schools with reciprocal rights and obligations. Therefore, school leaders are expected to play a greater role in leadership beyond school borders. Within the context of school empowerment or autonomy, decision-making by the school has been a consistent movement since the 1980s. The increase in autonomy required the school leaders to make budgetary changes, promote professional capacity development, rethink the design of school buildings, and consider many more aspects.

How might national and state policy frameworks (including curriculum and assessment, school quality and improvement) successfully engage and interact with key activities and characteristics of the school (including learning focus, structure, culture, and decision-making capacity)? What considerations must be taken when formulating policies of curriculum and implementation of policies within the classroom (class size, teaching approaches, and learning resources)? How does one optimize the capacity and work of school leaders to influence and promote effective learning? How might one be informed of the processes of influence beyond relying on interpretive and explanatory qualitative studies? Indeed, any attempt to design and carry out a comprehensive analysis of the ways in which leaders influence and promote successful outcomes through their decision-making will require specific methods and procedures beyond the traditional research methods (Leithwood & Levin, 2005). In particular, distributed leadership research stands to gain the most if relevant research methodologies were adopted that could be informative of the workings/actions of school leadership.

7.5 What Are the Alternatives to Current Social Science Methodologies for Educational Leadership?

As stated earlier, it is important to ensure that any alternative research methodologies proposed must adhere to the characteristic of disciplined inquiry. To further expand on this characteristic, Cronbach and Suppes stated that “Disciplined inquiry does not necessarily follow well-established, formal procedures. Some of the most excellent inquiry is free-ranging and speculative [...] trying what might seem to be a bizarre combination of ideas and procedures...” (Cronbach & Suppes, 1969, p. 16).

Drawing from the statement by Cronbach and Suppes, there are two other important points about disciplined inquiry that must be addressed here. Disciplined inquiry is not solely focussed on establishing facts. The methods of observation and inquiry are critical in defining which selection of facts of a phenomenon are found. Establishing facts can be done through a selection of observations and/or data collection methods. This point is not meant to raise the philosophical argument of positivism and post-positivism although it may be implied. Rather, from a pragmatic perspective, and to adhere to the characteristic of disciplined inquiry, one should be open to different types of observations and data collection methodologies, and thus different types of facts, as long as the definition of disciplined inquiry is adhered to.

To further support this view, it must be understood that the field of educational leadership is not a discipline by itself. As in any field of study, one should not be limited to a single discipline to dictate and direct the focus and forms of studies. Instead, procedures and perspectives of different disciplines, such as biology, chemistry, economics, geography, politics, anthropology, sociology, and others might bear on the research questions that can be investigated.

7.5.1 Brief Introduction to Complexity Science from an Educational Leadership Perspective

Complexity science appeared in the twentieth century in response to criticism of the inadequacy of the reductionist analytical thinking model in helping to understand systems and the intricacies of organizations. Complexity science does not refer to a single discipline; like in social science, a family of disciplines (psychology, sociology, economics, etc.) adopt methodologies to study society-related phenomena. Complexity science includes the disciplines of non-linear dynamical systems, networks, synergetics, and complex adaptive systems, and others.

The cornerstone concept of complexity science is the complex system. Complex systems have distinctive characteristics of self-organization, adaptive ability, emergent properties, non-linear interactions, and dynamic and network-like structures (Bar-Yam, 2003; Capra, 1996; Cilliers, 2001). By looking at the complex system of an organization, leadership should, consequently, be viewed in a different light. A complex system is a 'functional whole,' consisting of interdependent and variable parts. In other words, unlike a conventional system (e.g. an aircraft), the parts need not have fixed relationships, fixed behaviours, or fixed quantities. Thus, their individual functions may also be undefined in traditional terms. Despite the apparent tenuousness of this concept, these systems form the majority of our world, and include living organisms and social systems, along with many inorganic natural systems (e.g. rivers). The following is a brief introduction of key concepts of complexity science. These concepts are also the methodological assumptions for complexity science.

7.5.2 Emergence

Emergence is a key concept in understanding how different levels are linked in a system. In the case of leadership, it is about how influence happens at the individual, structural, and system levels. These different levels exist simultaneously, and one is not necessarily more important than the other, rather they are recognized as co-existing and linked.

Each level has different patterns and can be subjected to different kinds of theorization. Patterns at ‘higher’ levels can emerge in ways that are hard to predict at the ‘lower’ levels. The challenge (long-acknowledged in leadership research) is to understand how different levels interact and affect school outcome or school improvement. This question of the nature of ‘emergence’ has been framed in a variety of ways, including those of “macro-micro linkage,” “individual and society,” the “problem of order,” and “structure, action and structuration” (Giddens, 1984). In this paper, Giddens’ explanation of emergence as the relationship between the different levels through the “structure and agency” is adopted.

Giddens stated that the term “structure” referred generally to “rules and resources.” These properties make it possible for social practices to exist across time and space and that lends them ‘systemic’ form (Giddens, 1984, p. 17). Giddens referred to agents as groups or individuals who draw upon these structures to perform social actions through embedded memory, called memory traces. Memory traces are, thus, the vehicle by which social actions are carried out. Structure is also, however, the result of these social practices.

7.5.3 Non-linearity

Non-linearity refers to leadership effects or outcomes that are more complicated than being assigned to a single source or single chain of events. Influence and outcome are considered linear if one can attribute cause and effect. Non-linearity in leadership, however, means that the outcome is not proportional to the input and that the outcome does not conform to the principle of additivity, i.e. it may involve synergistic reactions, in which the whole is not equal to the sum of its parts.

One way to understand non-linearity is about how small events lead to large scale changes in systems. Within the natural sciences, the example often cited (or imagined) is that of a small disturbance in the atmosphere in one location, perhaps as small as the flapping of a butterfly’s wings, tipping the balance of other systems, leading ultimately to a storm on the other side of the globe (Capra, 1997).

7.5.4 Self-Organization

Self-organization happens naturally as a result of non-linear interactions among staff members in the school (Fontana & Ballati, 1999). As the word describes, there is no central authority guiding and imposing the interactions. Staff members adapt to changing goals and situations by adopting communication patterns that are not centrally controlled by an authority. In the process of working towards a goal (e.g. solving a leadership problem), self-organizing staff members tend to exhibit creativity and novelty as they have to quickly adapt and to find ways and means to solve the problem and achieve the goal.

This particular phenomenon is best observed in distributed leadership (Ng & Ho, 2012; Yuen, Chen, & Ng, 2015). As a result of interactions among members, the emergence of new patterns in conversation happens. This is an important aspect of self-organization. When there are no new patterns in conversations, there are no new ideas and no novel ways to solve problems. It must be noted that new patterns of conversation depend upon the responsiveness of its members towards each other and their awareness of each other's ideas and responses. As a result of the behaviour of interacting members, learning and adaptation, i.e. novel ways of solving problems emerge.

As stated earlier, complexity science is interdisciplinary and as such, there are multiple methods and ways to study complexity phenomena. It is nearly impossible to delve into these methodologies in a meaningful manner within the scope of one paper.

The intention with this paper is to propose alternative social science methodologies and analytical tools to perform educational leadership research. The following section will highlight one of the methods used in complexity science research that provides an alternative to the limitations identified in current research methodologies in educational leadership research.

7.6 Social Network Analysis as an Alternative to Normal Distribution and Linearity

Social Network Analysis (Scott, 2011; Wasserman & Faust, 1994) focuses on relational structures that characterize a network of people. These relational structures are represented by graphs of individuals and their social relations, and indices of structure, which analyze the network of social relationships on the basis of characteristics such as neighbourhood, density, centrality, cohesion, and others. The Social Network Analysis-method has been used to investigate educational issues, such as teacher professional networks (Baker-Doyle & Yoon, 2011; Penuel, Riel, Krause, & Frank, 2009), the spread of educational innovations (Frank, Zhao, & Borman, 2004), and peer influences on youth behaviour (Ennett et al., 2006). Table 7.4 provides examples of the types of data collected, and the analytical methods and analytical tools used in social network analysis.

In network analysis, indicators of centrality identify the most important vertices within a graph. Two separate measures of degree centrality, namely in-degree and out-degree, are used. In-degree is a count of the number of ties directed to the node (agent/individual) and out-degree is the number of ties that the node (agent/individual) directs to others. When ties are associated to positive aspects, such as friendship or collaboration, in-degree is often interpreted as a form of popularity and out-degree as a form of gregariousness.

For example, the study of Bird and colleagues (Bird, Gourley, Devanbu, Gertz, & Swaminathan, 2006) introduces social network analysis and the evidence of long-tailed distribution, which is a distinctive digression from the traditional social

Table 7.4 Social network data

Data:	Types and methods:
Types of data collected for Social Network Analysis	Social bonds (interpersonal ties, friendship, family networks)
	Organizational links (connection between residents and community organizations)
	Media connection (specific media that residents and organizations rely upon for news)
	Identify boundaries
	Clarify and design questions
	“Actually existing social relations”
	“Perceived relations”
	Dynamism: “Episodic” relations or “typical”/“long term” ties
Methods used to collect data for Social Network Analysis	Surveys
	Interviews
	Facebook, LinkedIn
	Data mining (internet, emails)
	Archival data
	Observations
Analytical tools for Social Network Analysis	Netlogo
	Netdraw
	UCINET
	NodeXL
	Gephi
	PAJEK
	SPAN
	STATNET

science study and the normal distribution associated with it. The evidence from social network measures in this research suggests that “developers who actually commit changes, play much more significant roles in the e-mail community than non-developers” (Bird et al., 2006, p. 142). What this conclusion alludes to is that knowledgeable and active developers who demonstrate their ability by actively responding and making changes (out-degree) based on feedback are more often contacted by e-mail queries from other users.

7.6.1 How Does Social Network Analysis Contribute to Educational Leadership Research?

The usefulness of social network analysis is reflected in a study (co-conducted by the author) on instructional leadership practices in primary schools in a centralized system where hierarchical structures are in place (Nguyen et al., 2017). It is noteworthy that the hierarchical structure’s inherent reliance on a ‘supreme leader’ is

greatly mitigated by the emergence of heterarchical elements. In brief, hierarchical structures, on the one hand, are vertical top-down control and reporting structures. Heterarchical structures, on the other hand, are horizontal. The findings revealed that at the teachers' and other key personnel's horizontal levels of hierarchy, spontaneous interactions and collaborations take place within a group and amongst groups of teachers. Through these horizontal professional interactions, individuals exert reciprocal influences on one another, with the minimal effects of authority power. In this structure, distributed instructional leadership appears to be deliberately practiced. Key personnel and teachers work in collaborative teams and are supported by organizational structures, initiated by the principals. This is where various instructional improvement programmes and strategies are initiated, implemented, and led by staff members. This would be highly impossible, if the principal practices were heavily based on hierarchical instructional leadership.

This study implies that decision-making on instructional improvement programmes is rigorously and actively practiced by teachers at the heterarchical level. Decision-making involves getting support for resources and approval from authorities over the teachers. In an organizational hierarchical structure, it would be the authority immediately above the teachers - the Head of Department, followed by the Vice Principal, and finally the Principal. Typically, such a reporting and resource seeking structure would be ineffective in creating instructional improvement programmes. If one was to redo the study and adopt social network analysis measures, how would the findings be presented? The figures below are hypothetically generated to provide a possible way to interpret hierarchical and heterarchical structures: Fig. 7.1 shows a social network representation, which provides an alternative way to represent hierarchy. The central (purple dot) represents the Principal, while the connected red dots to the Principal are the Head of Departments. The Head of Departments then oversees Subject Heads and finally teachers. Implying from our

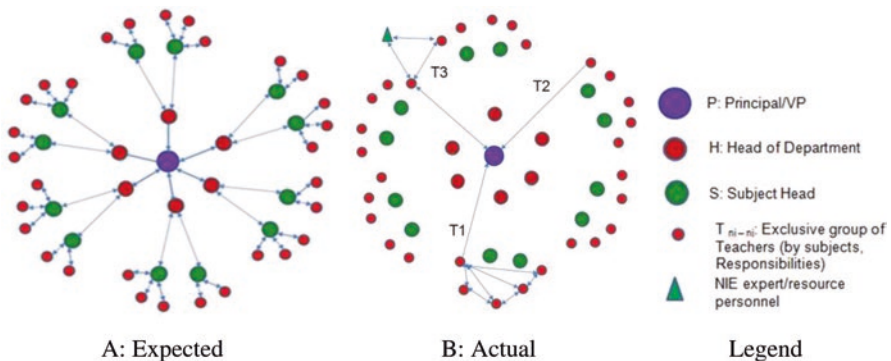


Fig. 7.1 Expected and actual reporting and decision-making pathways in managing teaching and learning

Note: In B, T1 = perceived authority for immediate action (e.g. allocation of resources, ability to act); T2 = perceived trust; T3 = pilot curriculum project

study, where heterarchical elements are exhibited, social network representation will most plausibly provide the means to represent the elements in Fig. 7.1.

What is immediately evident, is that the representation provides a more realistic way to look at social interactions involving decision-making. The connected dots among teachers could reveal who they interact most with. In addition, what would be most revealing is the emergence of how teachers in hybrid hierarchical and heterarchical structures make decisions. Specifically, the emergence of by-passing the constraints of a typical top-down hierarchical structure by directly getting support from centrality – the principal, who controls and provides resources and who also approves final decisions.

In summary, the discussion on one of the complexity science methodologies/social network analysis presents opportunities to reframe educational leadership research. It is now possible to ask research questions that are not bound by the constraints of current social science methodologies. Here are a number of questions using Social Network Analysis alone:

- What is the local (indigenous) knowledge base of instructional leadership and how does it emerge?
- How do different level leaders (Ministry of Education, Superintendents, Principals, etc.) shape the perception of curriculum policies in schools? (And – for specific local understanding – who are the influential personnel impacting curriculum and policy implementation?)
- Examination of ties among school departments that affect school improvement: What are the implications for long-term strategy processes for school improvement in light of the complex and adaptive nature of departments?
- What does engagement in decision-making look like?
- How do aspects of relations within the network: structural (pattern of interaction, face-to-face interaction), affective (benevolence and trust), and cognitive (mutual knowledge about each other's skills and knowledge, and shared systems of meaning) affect professional development and learning?
- Will an intended school vision/policy enjoy the desired positive reception among staff members?

7.7 Conclusion

This chapter contains the review that social science methodologies and analytical tools have been consistently and almost universally adopted in educational leadership research for the last three decades. This paper also highlights a number of limitations of current social science methodologies. The alternative complexity science research methodologies proposed are not merely alternative or novel ways of examining the problems or issues encountered. What is more valuable is that these

alternative methodologies bring with them their contrasting disciplinary roots and their corresponding (new) questions. The interest in the effects of educational leadership on school improvement can now be investigated by asking different research questions. One could, indeed, go deeper, wide-angle or zoom-in, and even make predictions by revisiting the basic question of “What do we wish to know about school improvement that we do not yet know enough about?”

By being open to alternative methodologies, one has nothing to lose but everything to gain in the scholastic pursuit of knowledge in the field of educational leadership and management. Researchers must avoid being educational leadership researchers who see the world merely from the perspectives that they have lived in; they should also avoid accepting these perspectives as the only perspectives without questions. The choice of research method or combination of methods affects the type of research questions asked (although, in practice, the questions are also often shaped by the researchers’ training and area of expertise). Ideally, one should not be constrained by methods before asking research questions. Research questions are the primary drivers of the quest for knowledge. This is the basis from which the most relevant methodologies are found that can answer research questions and provide researchers with the findings that can contribute to theory formation, knowledge building, and translation into practice. The author, therefore, proposes the following implications for practice and for research:

- Introduce complexity science (and also other disciplines) as additional graduate research courses. One can still tap on the transmission-form of knowledge transfer and supervisor-supervisee platform.
- Partner with established experts in the discipline of complexity science to leverage and speed up transfer of learning and research skills among educational leadership professors.
- Engage in epistemological and ontological discussions (including generalizability of findings) on complexity theory – to deepen our understanding of the advantages and limitations of complexity science disciplined inquiries.
- Expand educational leadership journals to accept findings and research that do not necessarily conform to social science methodologies alone.

Finally, reframing educational leadership research is an imperative in the light of diminishing researchable aspects due to the limitations of current methodologies. I, the author, want to reiterate that I do not advocate replacing existing social science methodologies. I acknowledge that social methodologies are still essential and vital. The full spectrum of social science research methodologies is needed to continue contributing to theory development in educational leadership and management. However, one also needs alternatives and complementary approaches to social science, such as complexity science methodologies for both theory development and theory building. The important thing to remember is that the questions come first and the methods follow.

References

- Andriani, P., & McKelvey, B. (2007). Beyond Gaussian averages: Redirecting international business and management research toward extreme events and power laws. *Journal of International Business Studies*, 38(7), 1212–1230.
- Baker-Doyle, K. J., & Yoon, S. A. (2011). In search of practitioner-based social capital: A social network analysis tool for understanding and facilitating teacher collaboration in a US-Based STEM professional Development program. *Professional Development in Education*, 37(1), 75–93.
- Bar-Yam, Y. (2003). Unifying principles in complex systems in converging technology (NBIC) for improving human performance. In M. C. Roco & W. S. Bainbridge (Eds.), *Converging technologies for improving human performance: Nanotechnology, biotechnology, information technology and cognitive science* (pp. 380–409). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Bass, B. M. (1997). Does the transactional–transformational leadership paradigm transcend organizational and national boundaries? *American Psychologist*, 52(2), 130.
- Bennett, N., Wise, C., Woods, P., & Harvey, J. A. (2003). *Distributed leadership*. Available at: www.ncsl.org.uk/media/3C4/A2/distributed-leadership-literature-review.pdf. Accessed 28 July 2006.
- Bhattacharjee, A. (2012). *Social science research: Principles, methods, and practices*. Textbook Collections, 3rd Global Text Project. Retrieved from http://scholarcommons.usf.edu/oa_textbooks/3/
- Bird, C., Gourley, A., Devanbu, P., Gertz, M., & Swaminathan, A. (2006, May). Mining email social networks. In *Proceedings of the 2006 international workshop on mining software repositories* (pp. 137–143). New York, NY: ACM.
- Boyan, N. J. (1981). Follow the leader: Commentary on research in educational administration. *Educational Researcher*, 10(2), 6–13. 21.
- Capra, F. (1996). *The web of life: A new scientific understanding of living systems* (1st ed.). New York, NY: Anchor.
- Capra, F. (1997). *The name above the title: An autobiography*. New York, NY: Da Capo Press.
- Cilliers, P. (2001). Boundaries, hierarchies and networks in complex systems. *International Journal of Innovation Management*, 5(2), 135–147.
- Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.
- Cronbach, L. J., & Suppes, P. (1969). *Research for tomorrow's schools: Disciplined inquiry for education*. New York, NY: Macmillan.
- Cziko, G. (1989). Unpredictability and indeterminism in human behavior: Arguments and implication for educational research. *Educational Researcher*, 18, 17–25.
- Darling-Hammond, L. (2010). Teacher education and the American future. *Journal of Teacher Education*, 61(1–2), 35–47.
- Day, C., Sammons, P., Leithwood, K., Hopkins, D., Harris, A., Gu, Q., & Brown, E. (2010). *Ten strong claims about successful school leadership*. Nottingham, UK: The National College for School Leadership.
- Dimmock, C. (2000). Globalisation and societal culture: Redefining schooling and school leadership in the twenty-first century. *Compare*, 30(3), 1–6.
- Dimmock, C., & Walker, A. (2005). *Educational leadership: Culture and diversity*. London, UK: Sage.
- Doll, W. E., Jr. (1989). Foundations for a post-modern curriculum. *Journal of Curriculum Studies*, 21(3), 243–253.
- Ennett, S. T., Bauman, K. E., Hussong, A., Faris, R., Foshee, V. A., Cai, L., & DuRant, R. H. (2006). The peer context of adolescent substance use: Findings from social network analysis. *Journal of Research on Adolescence*, 16(2), 159–186.

- Fontana, W., & Ballati, S. (1999). Complexity. *Complexity*, 4, 14–16. [https://doi.org/10.1002/\(SIC\)1099-0526\(199901/02\)4:3<14:AID-CPLX3>3.0.CO;2-O](https://doi.org/10.1002/(SIC)1099-0526(199901/02)4:3<14:AID-CPLX3>3.0.CO;2-O)
- Frank, K. A., Zhao, Y., & Borman, K. (2004). Social capital and the diffusion of innovations within organizations: The case of computer technology in schools. *Sociology of Education*, 77(2), 148–171.
- Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Oakland, CA: University of California Press.
- Glatter, R., & Kydd, L. (2003). Best practice in educational leadership and management: Can we identify it and learn from it? *Educational Management & Administration*, 31(3), 231–243.
- Goldring, E., & Greenfield, W. (2002). Understanding the evolving concept of leadership to education: Roles, expectations, and dilemmas. *Yearbook of the National Society for the Study of Education*, 101(1), 1–19.
- Griffiths, D. E. (1959). *Administrative theory*. New York, NY: Appleton-Century-Crofts.
- Griffiths, D. E. (1979). Intellectual turmoil in educational administration. *Educational Administration Quarterly*, 15(3), 43–65.
- Gronn, P. (2002). Distributed leadership as a unit of analysis. *Leadership Quarterly*, 13, 423–451.
- Gronn, P. (2009). From distributed to hybrid leadership practice. In *Distributed leadership* (pp. 197–217). Springer, Dordrecht.
- Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education*, 33(3), 329–351.
- Hallinger, P. (2011). Leadership for learning: Lessons from 40 years of empirical research. *Journal of Educational Administration*, 49(2), 125–142.
- Hallinger, P. (2013). A conceptual framework for reviews of research in educational leadership and management. *Journal of Educational Administration*, 51(2), 126–149.
- Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980–1995. *Educational Administration Quarterly*, 32(1), 5–44.
- Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *The Elementary School Journal*, 86(2), 217–247.
- Hallinger, P., & Murphy, J. F. (1986). The social context of effective schools. *American Journal of Education*, 328–355.
- Hallinger, P., Walker, A. D., & Bajunid, I. A. (2005). Educational leadership in East Asia: Implications for education in a global society. *UCEA Review*, 1, 1–4.
- Hargreaves, A. P., & Shirley, D. L. (Eds.). (2009). *The fourth way: The inspiring future for educational change*. Thousand Oaks, CA: Corwin Press.
- Harris, A. (2005). OP-ED. *Journal of Curriculum Studies*, 37(3), 255–265.
- Howell, J. M., & Avolio, B. J. (1993). Transformational leadership, transactional leadership, focus of control, and support for innovation: Key predictors of consolidated business-unit performance. *Journal of Applied Psychology*, 78(6), 891.
- Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics*, 18(2), 141–165.
- Lee, M., Hallinger, P., & Walker, A. (2012). A distributed perspective on instructional leadership in International Baccalaureate (IB) schools. *Educational Administration Quarterly*, 48(4), 664–698.
- Leithwood, K., & Day, C. (2007). *Successful school leadership in times of change*. Toronto, ON: Springer.
- Leithwood, K., & Duke, D. (1999). A century's quest to understand school leadership. In J. Murphy & K. Louis (Eds.), *Handbook of research on educational administration* (pp. 45–72). San Francisco, CA: Jossey-Bass.
- Leithwood, K., & Jantzi, D. (1999). Transformational school leadership effects: A replication. *School Effectiveness and School Improvement*, 10(4), 451–479.
- Leithwood, K., & Jantzi, D. (2000). The effects of transformational leadership on organizational conditions and student engagement with school. *Journal of Educational Administration*, 38(2), 112–129.

- Leithwood, K., & Levin, B. (2005). *Understanding leadership effects on pupil learning*. Toronto, UK: UK Department of Skills and Education.
- Leithwood, K., Tomlinson, D., & Genge, M. (1996). Transformational school leadership. In K. Leithwood, J. Chapman, D. Corson, P. Hallinger, & A. Hart (Eds.), *International handbook of educational leadership and administration* (pp. 785–840). Dordrecht, The Netherlands: Springer.
- Luo, W., Hogan, D. J., Yeung, A. S., Sheng, Y. Z., & Aye, K. M. (2014). Attributional beliefs of Singapore students: Relations to self-construal, competence and achievement goals. *Educational Psychology, 34*(2), 154–170.
- MacBeath, J., & Cheng, Y. C. (2008). *Leadership for learning. International perspective* (pp. 17–40). Rotterdam, The Netherlands: Sense Publishers.
- Meyer, M. J., & Macmillan, R. B. (2001). The principal's role in transition: Instructional leadership ain't what it used to be. *International Electronic Journal for Leadership in Learning, 5*(13), 1–14.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156.
- Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics, 1*(2), 226–251.
- Mulford, B., & Silins, H. (2003). Leadership for organisational learning and improved student outcomes – What do we know? *Cambridge Journal of Education, 33*(2), 175–195.
- Murphy, J. (2015). Forces shaping schooling and school leadership. *Journal of School Leadership, 25*(6), 1064.
- Ng, F. S. D., & Ho, J. M. (2012). How leadership for an ICT reform is distributed within a school. *International Journal of Educational Management, 26*(6), 529–549.
- Nguyen, D. H., Ng, D., & Yap, P. S. (2017). Instructional leadership structure in Singapore: A co-existence of hierarchy and heterarchy. *Journal of Educational Administration, 55*(2), 147–167.
- Nicolis, G., Prigogine, I., & Nocolis, G. (1989). *Exploring complexity*. New York, NY: Freeman.
- Penuel, W., Riel, M., Krause, A., & Frank, K. (2009). Analyzing teachers' professional interactions in a school as social capital: A social network approach. *The Teachers College Record, 111*(1), 124–163.
- Ritchie, J., Lewis, J., Nicholls, C. M., & Ormston, R. (Eds.). (2013). *Qualitative research practice: A guide for social science students and researchers*. London, UK: Sage.
- Robinson, V. M. (2010). From instructional leadership to leadership capabilities: Empirical findings and methodological challenges. *Leadership and Policy in Schools, 9*(1), 1–26.
- Robinson, V., Lloyd, C., & Rowe, K. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly, 44*(5), 564–588.
- Scott, J. (2011). Social network analysis: Developments, advances, and prospects. *Social Network Analysis and Mining, 1*(1), 21–26.
- Sheppard, B. (1996). Exploring the transformational nature of instructional leadership. *Alberta Journal of Educational Research, 42*(4), 325–344.
- Southworth, G. (2002). Instructional leadership in schools: Reflections and empirical evidence. *School Leadership and Management, 22*(1), 73–92.
- Spillane, J. P. (2005, June). Distributed leadership. *The Educational Forum, 69*(2), 143–150.
- Spillane, J. P., Diamond, J. B., & Jita, L. (2003). Leading instruction: The distribution of leadership for instruction. *Journal of Curriculum Studies, 35*(5), 533–543.
- Spillane, J. P., Halverson, R., & Diamond, J. B. (2001). Investigating school leadership practice: A distributed perspective. *Educational Researcher, 30*(3), 23–28.
- Spillane, J. P., Halverson, R., & Diamond, J. B. (2004). Towards a theory of leadership practice: A distributed perspective. *Journal of Curriculum Studies, 36*(1), 3–34.
- Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., & Deller Jr, J. R. (2002, September). Approaches to language identification using Gaussian mixture models

- and shifted Delta Cepstral features. In *Proceedings of the international conference on spoken language processing*, Denver, CO, pp. 89–92.
- Von Bertalanffy, L. (1968). *Organismic psychology and systems theory*. Worcester, MA: Clark University Press.
- Walberg, H. J., Strykowski, B. F., Rovai, E., & Hung, S. S. (1984). Exceptional performance. *Review of Educational Research*, 54(1), 87–112.
- Walker, A., & Dimmock, C. (2000). Insights into educational administration: The need for a cross-cultural comparative perspective. *Asia-Pacific Journal of Education*, 20(2), 11–22.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). New York, NY: Cambridge University Press.
- Witziers, B., Bosker, R., & Kruger, M. (2003). Educational leadership and student achievement: The elusive search for an association. *Educational Administration Quarterly*, 34(3), 398–425.
- Woods, P. A. (2004). Democratic leadership: drawing distinctions with distributed leadership. *International journal of Leadership in Education*, 7(1), 3–26.
- Yuen, J. H. P., Chen, D. T. V., & Ng, D. (2015). Distributed leadership through the lens of activity theory. *Educational Management Administration & Leadership*, 44(5), 814–836.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 8

The Structure of Leadership Language: Rhetorical and Linguistic Methods for Studying School Improvement



Rebecca Lowenhaupt

8.1 Introduction

As the field of educational leadership evolves, there has been an increased focus on school-level leaders as architects and implementers of reform efforts. Research has established the importance of these local leaders, emphasizing the ways school leaders can create the conditions and capacity for enacting change (Spillane, 2012). While this research has focused on leadership *actions*, earlier work reminds us of the often overlooked yet crucial actions that occur in the form of leadership *talk*, one of the most prevalent and influential forms of leadership practice (Gronn, 1983). Indeed, school leaders use language both to describe and to enact practice, as talk is often the medium through which key actions occur within schools (Lowenhaupt, 2014).

Building theory about the language of school leadership, this chapter considers the frameworks and methodologies used to study the everyday communication strategies leaders use. In so doing, I aim to describe both why and how one might study principal talk. As illustrated through various analyses of discourse in organizational studies (Alvesson & Kärreman, 2000; Suddaby & Greenwood, 2005), language is a fundamental feature of social organizations (Gee, 1999; Heracleous & Barrett, 2001), and the leadership of those organizations (Gronn, 1983; Mehan, 1983). I argue that understanding the role of leadership in school improvement requires deeper study of the form and content of language used to enact reform.

Framing language as action, this chapter explores the methodological implications of attending to leadership language. I consider how research about the ways leaders use language in their daily practice might contribute important insights into how leadership shapes school improvement. Understanding how language is used as

R. Lowenhaupt (✉)
Boston College, Newton, MA, USA
e-mail: rebecca.lowenhaupt@bc.edu

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research, Accountability and Educational Improvement*, https://doi.org/10.1007/978-3-030-69345-9_8

137

a tool for enacting reform can shed light on the microprocesses of school improvement.

After first considering the role of language in principal practice, I then discuss the methods associated with linguistic analyses and explore how those methods might be used in the study of school leadership. I then share examples from my previous work, before concluding with a discussion of implications for future work. Overall, this chapter aims to demonstrate how language is a crucial feature of leadership practice, and one, which must not be neglected in research about school effectiveness and improvement.

8.2 School Leadership and School Improvement

In the last few decades, policymakers at federal, state, and district levels have increasingly looked to school principals to implement school-level reforms (Darling--Hammond, LaPointe, Meyerson, & Orr, 2007; Horng, Klasik, & Loeb, 2010; Spillane & Lee, 2014). Improvement efforts focused on standardizing curricula, enacting accountability measures and developing teacher evaluation systems all depend on the work of principals to implement them in their schools (Kraft & Gilmour, 2016; Lowenhaupt & McNeill, 2019). While previous conceptions of the principal role focused primarily on managerial tasks, along with buffering teacher autonomy (Deal & Celotti, 1980; Firestone, 1985; Firestone & Wilson, 1985), principals are now asked to lead efforts to develop professional communities, support instructional improvement, and bridge classrooms, family, and community (Lowenhaupt, 2014; Rallis & Goldring, 2000).

In response, a focus on principal practice has emerged in recent research with efforts to understand how specific practices influence school effectiveness (Camburn, Spillane, & Sebastian, 2010; Grissom & Loeb, 2011; Horng et al., 2010; Klar & Brewer, 2013). Although many of these studies are quantitative, various qualitative studies also contribute to our understanding of school principals as they navigate a range of responsibilities. In the tradition of longstanding in-depth work about the role (Dillard, 1995; Gronn, 1983; Peterson, 1977; Wolcott, 1973), these studies develop portraits about the daily work and practice of school leaders in an increasingly complex reform context (Browne-Ferrigno, 2003; Khalifa, 2012; Spillane & Lee, 2014; Lowenhaupt & McNeill, 2019; Spillane & Lowenhaupt, 2019). Employing a range of methodologies, from surveys and administrative logs to ethnographic observations and interviews, these studies highlight the various roles and complexities these school-level leaders navigate in the context of improvement efforts.

Importantly, this research elaborates on the diversity of tasks principals engage in throughout their days, as they enact their various responsibilities. As instructional leaders (Goldring, Huff, May, & Camburn, 2008; Hallinger, 2005), meaning-makers (Bolman & Deal, 2003; Dillard, 1995; Peterson, 1977), coalition-builders (Lortie, 2009), managers (Goldring et al., 2008; Horng et al., 2010), and community leaders

(Dillard, 1995; Khalifa, 2012; Peterson, 1977), they work with stakeholders both within and outside of their schools to ensure school effectiveness. As such, interactions are a crucial part of their work via building strong relationships, bringing stakeholders together, and mediating conflict (Peterson & Kelley, 2002; Rallis & Goldring, 2000). All these responsibilities depend on the use of language to communicate a vision, negotiate competing demands, and promote reforms. Yet too often, research treats language as the medium for action without attending to the language as an integral part of the practice itself (Lowenhaupt, 2014).

8.3 Leadership Language as Action

Although a robust body of research has emerged related to these new school leadership practices, only a handful of scholars have turned their attention explicitly to the language used to enact them. These scholars have argued for the need for further research about the discourse of leadership (Lowenhaupt, 2014; Riehl, 2000). Recognizing that “talk is the work” (Gronn, 1983), a handful of scholars have employed discourse and linguistic analytic methodologies to explore how leaders use language as practice.

Some researchers have focused on the linguistic strategies principals develop to persuade teachers to shift their practice (Gronn, 1983; Lowenhaupt, 2014), while others have explored how language is central to the symbolic meaning-making principals engage in to develop school culture (Deal & Peterson, 1999). By shaping communication, spoken or written, formal or informal, to argue for particular outcomes, principals draw on a range of rhetorical and linguistic repertoires to enact their leadership. As such, language ought to be viewed as a practice, which leaders can and often do purposefully and strategically employ in relation to others.

Importantly, this leadership language cannot be viewed as one-directional or limited to an individual leader. Theories of distributed leadership have emphasized that leadership is shared across individuals and in relationship between leaders and followers (Leithwood, Harris, & Hopkins, 2008; Spillane, 2012). In order to understand how language functions within the context of interactions, scholars need to move beyond the language of individual leaders to study the negotiations and discussion that occur in conversations among various stakeholders (Gronn, 1983; Mehan, 1983; Riehl, 1998). An important focus for these interaction analyses is the linguistic processes that play out in meetings and the ways in which language influences and informs the change process among administrators and teachers (e.g. Riehl, 1998). In another example, Mehan (1983) looked at the administrative process of Special Education identification and the form and content of discourse in meetings among administrators, staff, and families. In both cases, these studies identified features of language that influenced outcomes for students and educators. Taken together, these various studies point to the need for further study of everyday language that considers the levers of change particular leaders employ through their talk.

While some of these interactions are public, high-stakes forms of talk, it is important to highlight that leadership language occurs in both informal and formal settings. Although principals are called on to give speeches, write public statements, and interact during public forums, they also engage in conversation throughout their day-to-day work. This prior scholarship reminds us that this talk, particularly in the context of reform efforts, is never neutral. Indeed, these various interactions work as a form of persuasion with political implications, as well as implications for school effectiveness.

Turning the lens on the linguistic form and content of these interactions reminds us that language both describes and creates actions. As such, language is both a means for enacting practice, as well as a practice in and of itself. Empirical study of leadership language requires discourse analyses focused on both the form and content of that language in distinct contexts in order to uncover exactly how principals use language toward school effectiveness (Riehl, 2000). A linguistic turn in the study of school leadership requires a shift in methodologies to uncover the ways in which language manifests itself as action. I turn to a discussion on methodology next.

8.4 Language in Organizations

Educational leadership is not the only field to seek a linguistic turn in social science research. Across the social sciences and within education, various forms of discourse analyses have developed as a methodology for interpreting language practices within complex socio-cultural contexts (Gee, 1999). In the field of organizational studies, scholars have also drawn on studies of discourse to understand how everyday language shapes the nature of those organizations (Alvesson & Kärreman, 2000; Heracleous & Barrett, 2001; Watson, 1995). Across these fields, research has drawn attention to the ways in which various forms of language are used to, “continually and actively build and rebuild our world” (Gee, 1999, p. 11).

Language in organizations takes on many forms. In addition to formal written policies, which instantiate structures and systems, language also manifests itself through informal everyday interactions which constitute the social nature of organizations (Alvesson & Kärreman, 2000; Hallett, Harger, & Eder, 2009). During meetings, hallway conversations, and gossip in the workplace, people use language to share opinions, interpret realities, and shape practice (Hallett et al., 2009). For school leaders, talk is a central way by which formal policies are implemented in schools (Lowenhaupt, Spillane, & Hallett, 2016). The proliferation of digital communications through email, social media, and text messaging have further expanded the linguistic repertoires of the workplace.

Taken together, this complex ecosystem of language use within organizations provides ample fodder to researchers focused on investigating how language shapes leadership practice in schools. Drawing on the tools of discourse analysis, researchers might examine how the form and content of particular features of leadership

language influence improvement. In the context of school improvement, where leaders work to enact deep reform, I argue that rhetoric, or the language of persuasion, is a particularly fruitful area of inquiry, as I discuss in more detail next.

8.5 Rhetorical Analyses

To examine the everyday leadership language that is used in school improvement, rhetorical analysis provides the methodological tools to understand how persuasion works in the context of school improvement. Within a reform context, school leaders must establish the rationale for change and engage both staff and community members in new activities. One key mechanism for this is talk, and more specifically, persuasion. For leaders within these organizations, persuasion is a key, yet often implicit, feature of the social dynamics that lead to (or hinder) organizational change (Suddaby & Greenwood, 2005). Rhetoric is defined as the linguistic features of persuasion (Corbett & Connors, 1999). Within organizations, the role of rhetoric is one of the least well understood forms of coordination and control (Stone, 1997).

Recent work in organizational studies has drawn on rhetorical analyses to develop an understanding of how linguistic patterns influence the structure of organizations and lead to institutional change (Alvesson & Kärreman, 2000; Brown, Ainsworth, & Grant, 2012; Mouton, Just, & Gabrielsen, 2012; Suddaby & Greenwood, 2005). Similarly, the field of educational leadership might develop methods for rhetorical analyses to explore one form of language particularly relevant to unpacking leadership practice for school improvement.

The study of rhetoric focuses on both the form and content of language to reveal the linguistic structures of persuasion. Defined as the language used to persuade an audience, classical rhetoric continues to undergird the structure of our everyday language today (Corbett & Connors, 1999). As a method used in organizational studies, rhetorical analyses uncover implicit structures of persuasive language to demonstrate the, “recurrent patterns of interests, goals, and shared assumptions that become embedded in persuasive texts” (Suddaby & Greenwood, 2005, p. 49). While some focus on written text, others analyze spoken language to examine everyday interactions integral to the function of organizations (Gill & Whedbee, 1997).

Rhetorical analyses rely on strategies of textual analysis to explore linguistic features and patterns. As with other types of thematic qualitative analyses, systematic coding of text allows for the identification of forms and features of rhetoric. Working with transcripts, written communications, or other text, one can make use of various qualitative coding software to identify, select, and analyze particular linguistic segments that play a role in persuasion. By looking systematically at particular elements of language, one can uncover the underlying patterns and features of rhetoric. In particular, coding focused on audience, form, and content comprises analyses of rhetorical features.

One fundamental aspect of rhetoric is an emphasis on audience (Corbett & Connors, 1999). Drawing on various rhetorical forms, the speaker shapes rhetoric to

influence specific audience members in particular ways. Although not always purposeful or strategic, speakers draw on various linguistic forms to persuade depending on the particular orientation of the audience (Corbett & Connors, 1999). In terms of school leadership, this means using distinct rhetorical arguments depending on the various stakeholders involved, whether families, staff, community members, or students. Accordingly, rhetorical analyses take into consideration the social dynamics of the speaker-audience relationship and explore differences in argumentation as the audience shifts. This emphasis is in line with distributed leadership theory, which urges researchers to look at the interactions among leaders and followers as an interactive, socially constructed perspective on leadership (Spillane, 2012). Bringing rhetoric and leadership together, then, encourages research that looks at the language of interactions among leaders and various stakeholders. Taking this into account, textual analysis can attend to differences among stakeholders and compare varying uses of rhetoric based on audience.

In addition to a focus on audience, classical rhetoric also places form at the heart of understanding persuasion. Rhetorical analysis often begins with an examination of three primary forms central to argumentation, namely *logos*, *ethos*, and *pathos* (Corbett & Connors, 1999). The rational appeal, *logos*, uses reasons and justifications as an appeal to an audience's intellect (Suddaby & Greenwood, 2005). This form of appeal may vary by audience, as what seems logical to one group may be adapted for another group. Regardless, the key basis of persuasion for *logos* is reasoning and logic. In the context of school improvement, leaders might provide rational arguments for change and emphasize the need for improvement based on evidence, such as student achievement. Another form of argument, *ethos*, draws on the underlying ethics or values held by a particular organization. As such, the speaker makes an ethical claim that the argument aligns well with the values and orientation of the audience. While such appeals are often implicit throughout the interaction, rhetoric is considered *ethos* when it occurs as a specific and explicit argument used to establish the relatability and legitimacy of the speaker in espousing similar ethical values (Corbett & Connors, 1999). Often, leaders rely on the *ethos* of care for students or a sense of social obligation to motivate improvement efforts. Finally, the emotional appeal, or *pathos*, draws on the affective side of the argument to persuade. Arguably the most complex form, *pathos* is considered an appeal to the imagination and often takes the form of evocative storytelling or sharing emotionally charged examples, an appeal to the heartstrings (Corbett & Connors, 1999). School leaders might share anecdotes about student successes or hardships to motivate and inspire improvement.

While there are other structural features identified in classical rhetoric, these three forms are embedded throughout persuasive language and provide a meaningful frame for rhetorical analyses. By considering the rhetorical form for each segment of text and exploring the pattern of use across multiple forms, one can uncover the underlying structure of persuasion leaders use to try to convince others to enact improvement. Importantly, forms may be interwoven or occur independently throughout both formal and informal persuasion. Sometimes, these forms may co-occur, as leaders simultaneously draw on multiple forms of appeal. The ways in

which they are used and the relative affordances of each vary according to the speaker-audience relationship and the context of the argument (Aristotle, 1992).

While both audience and form are crucial areas of focus for rhetorical analyses, the language of persuasion also relies on content specific to the argument at hand. In the case of school leaders, that content is developed based on the particular initiatives and reforms leaders seek to enact for school improvement. Yet, the content also builds on longstanding values and professional norms in the field of education, as well as the particular school and community cultures in which leaders work. In other words, the implementation of new policies does not occur in a vacuum, but rather builds on and intersects with existing practices, beliefs, and knowledge (Spillane, 2012). As such, for persuasion to work, leaders must take up and navigate these existing socio-cultural aspects of their context. The content of rhetoric can serve to illuminate how new initiatives link to current context (Lowenhaupt et al., 2016). In other words, rhetorical content can construct a bridge between longstanding ways of thinking about the meaning and purpose of the work and new practices for school improvement.

Bringing together these three elements of audience, form, and content, rhetorical analyses can help identify meaningful patterns of persuasion and reveal how leadership language shapes school improvement. To conduct such analyses, identifying meaningful instances of language use and transforming it into transcripts or text can support a systematic coding process. Audio or video-recording, email communications, or other written artifacts can thus become data sources. Meeting transcripts are a particularly promising source, as leaders must often present the case of their improvement efforts to various audiences. By creating a coding structure and applying a systematic process through a qualitative coding software, such as Nvivo or Dedoose, researchers can enact rigorous rhetorical analyses. Using a combination of deductive and inductive approaches can make visible both the inherent linguistic structure and the shape of the argument. For example, applying a priori codes for logos, ethos, and pathos reveals rhetorical forms and sequences. At the same time, emergent, thematic coding for content can reveal the key arguments leaders use to persuade.

The linguistic turn in organizational studies provides fruitful lessons for the study of school improvement, and more specifically, the role of leadership in enacting reform. Drawing on various tools of discourse analyses, a focus on language can provide opportunities to learn about and subsequently shape the discursive practice of leadership in schools. Rhetorical analyses provide one possible framework with which to develop research methods for examining the linguistic features of leadership. Given the need for deeper understanding of how principals use language to both describe and enact reforms, I argue that the study of rhetoric holds substantial promise as a methodological approach to understanding leadership practice, particularly within the context of school improvement and change. To illustrate the potential of this approach, I next turn to an example of one study, which applied classical rhetoric to the analysis of leadership language.

8.6 Rhetorical Form and Principal Talk: An Example

Through a series of rhetorical analyses of one principal's language in various meetings during a year of school improvement, my collaborators and I investigated the rhetorical forms and content used to enact substantial reform in one urban public school (See Lowenhaupt, 2014; Lowenhaupt et al., 2016, for the complete studies). Working with data from a larger study of school reform led by Dr. James Spillane at Northwestern University and along with Dr. Timothy Hallett at Indiana University, who conducted the initial fieldwork, our team analyzed the rhetoric used by Mrs. Kox, an urban elementary school principal, to advocate for reform.

As a new principal, she was charged with implementing accountability measures focused on increasing student achievement. With support from the district, she increased classroom visits, encouraged standardization across classrooms, and conducted an audit of instruction focused on achievement measures. As she implemented these reforms, researchers observed and recorded many of her interactions with teachers, families, and other administrators as part of an in-depth ethnographic case study.

8.6.1 *Methods*

Analyzing 14 transcripts from two types of administrative meetings, we documented the microprocesses of organizational talk in meetings, key sites for organizational work (Riehl, 1998). External stakeholders were engaged through School Council meetings, where locally elected community members discussed initiatives with the principal. Empowered to represent the best interests of the community and overseeing the management of the school, this group was also responsible for evaluating the principal. Non-elected members of the community were also often present at these public meetings, where recent initiatives, policy reforms, and school change were discussed. Internal stakeholders participated in similar conversations in closed Leadership Team meetings, where select teachers and staff engaged in conversations about how to enact reforms.

We engaged in a series of textual analyses to surface the form and content of Mrs. Kox's rhetoric and explored how these aspects of rhetoric differed by audience. Taken together, these analyses presented insight into how to put into practice a rhetorical analysis of principal talk, as well as some considerations for this approach. Using qualitative coding software, Nvivo, we initiated the analysis by creating discrete segments of principal rhetoric ranging from a few words to full sentences (Suddaby & Greenwood, 2005). Decisions about where a particular 'utterance' began and ended were made with rhetorical form in mind, but drew on the context of the meeting as well (Gee, 1999; Goffman, 1981). For example, in one meeting, Mrs. Kox stated, "We need to define the curriculum because there is a need for consistency throughout the grades." In this case, the utterance was defined as a

complete sentence because it constituted a rhetorical unit with a claim, the need to define the curriculum, along with a rationale for that claim, the need for consistency. In other instances, one sentence consisted of multiple claims, in which case we coded clauses within sentences as discrete utterances. And in other instances, although rare, we coded multiple sentences as one utterance if it consisted of one rhetorical idea.

In this way, even at the early stages of analyses, the rhetorical framework influenced the process. Recognizing the importance of counter-argument as an influence on the persuasive process (Goffman, 1981; Symon, 2005), the analytic decision to focus exclusively on principal talk was primarily logistical, based on the need to focus on a manageable subset of utterances for analysis. Ultimately and across all 14 meeting transcripts, 650 utterances were coded as instances of principal rhetoric. We accounted for interaction through iterative analyses that looked at particular utterances in the broader context of discourse as well.

Once these utterances were identified, we worked as a research team on an iterative coding process. We conducted four distinct stages of analyses to examine form, content, audience, and sequences. During the first stage of analysis, two researchers independently coded approximately 20% of the total set of utterances according to a deductive, closed coding scheme of the three rhetorical forms, *logos*, *ethos*, and *pathos* (Corbett & Connors, 1999). We also employed a code for 'other' that took into account utterances that were difficult to categorize and which we ultimately determined to fit within one of the three forms. Importantly, we did allow for coding in multiple categories. After calculating interrater reliability for each code, we then engaged in an arbitration process, discussing our rationale on how we coded each utterance and resolving any disagreements. This process led to refining definitions of these forms, identifying examples of particular forms, and creating a coding manual that clearly explicated these features of each code (See Table 8.1). We then applied the coding scheme to the remaining utterances.

A second stage of analysis aimed to identify the content of the arguments through an inductive, open coding process within each form. This second iteration yielded content-based codes that described the general themes that were treated with the various rhetorical forms. In this way, we aimed to capture both what the principal discussed through rhetoric, as well as to explore the deeper discourses she tapped into through her persuasive language (Alvesson & Kärreman, 2000; Gee, 1999). For example, her use of *ethos* tended to rely on either an effort to assert her own legitimacy to teachers by referring to her prior experiences as an educator or an appeal to the ethical obligation of doing 'what's best for kids'. This appeal to serving children is a longstanding, professional commitment among educators and seeks to persuade others by reminding them of this commitment. During this stage of analysis, we employed a similar, collaborative process, while working together to determine an initial set of thematic codes, applied and refined them through arbitration, and ultimately developed a set of sub-codes within each form, as depicted in Table 8.1.

Once the entire set of utterances was coded for form and sub-coded for content, we embarked on a third stage of analyses to explore the underlying structure of principal rhetoric as it related to audience. We used inferential statistics, specifically

Table 8.1 Coding structure

Form	Definition	Examples	Content subcodes
Logos: rational appeal	Through the use of justifications, examples, and evidence, the rational appeal attempts to persuade through the use of (or appearance of) logic (Corbett & Connors, 1999).	“We need to define the curriculum because there is a need for consistency throughout the grades.”	Professional knowledge
		“There are always going to be some standards and there will always be some guidelines.”	Common sense Appeal to authority
Ethos: ethical appeal	By which the speaker convinces the audience by his or her words that he/she is of high moral character. The ethical appeal, “must display a respect for the commonly acknowledged virtues and an adamant integrity” (Corbett & Connors, 1999, p.73)	“When I came to this school, I established a guideline.”	Morality related to children
		“The bottom line here is that we’re providing services to the children.”	Legitimacy as a leader
Pathos: emotional appeal	The emotional appeal persuades by engaging the emotions of the audience, an appeal to the imagination through illustrative stories and the use of exaggerated, emotional language.	“It’s really marvellous...there’s a lot of wonderful things happening in the school.”	Evoking pity
			Showing empathy
			Story
			Humor
			Enthusiasm

chi-square analyses, to compare findings by audience by comparing Kox’s rhetoric across meeting types. Taken together, these three stages of analysis facilitated both the study of the form and content of a principal’s use of rhetoric, as well as the interpretation of how this rhetoric varied by audience.

In a fourth follow-up analysis, we investigated what emerged as an important feature of principal talk, the linking of multiple utterances working in concert to create an integrated, bridging form of persuasion we called ‘accountability talk’ (Lowenhaupt et al., 2016). Through analysis of rhetorical sequences, we demonstrated how Mrs. Kox relied on multiple forms together, primarily logos but linking logos with ethos and pathos, to bridge her new initiatives and their rationale with longstanding commitments in the field. In this analysis of sequences, we moved between discrete utterances, groups of utterances, and the broader meeting context to identify how this accountability talk was constructed. At all stages of this process, we articulated and followed a set of systematic steps which allowed us to uncover the underlying structures that undergirded the persuasive language one principal used in the reform context.

8.6.2 Findings

Findings from these analyses demonstrated that the principal used multiple forms of rhetoric to link accountability initiatives to existing norms, relying primarily on rational logics (logos), but also incorporating ethical (ethos) and emotional (pathos) arguments to solicit support for reforms (Lowenhaupt, 2014). Her reliance on logos illustrated the importance of reason and logic, but this was not enough to persuade. At the same time that her improvement efforts centered on logos, she also drew on ethical and emotional appeals, particularly with teachers, who were most directly impacted by her initiatives. Further analyses illustrated how these forms were woven together into rhetorical sequences that served to integrate longstanding norms with emerging policy pressures into a type of speech we termed, “accountability talk” (Lowenhaupt et al., 2016).

Focusing on rhetorical structure not only reveals how language is used to persuade others to engage in school improvement, but also can play an active, key role in improvement efforts. In the example presented above, the principal relied on rhetoric to promote support for aspects of improvement, such as accountability. Her use of rhetorical form established certain ideas as logical and asserted the importance of logic in the design of school improvement. She anchored this in treasured values of schooling by appealing to a sense of social obligation. The very structure of her rhetoric reminds both internal and external stakeholders that logic alone is not the motivation for improvement. As such, rhetoric can be viewed as a tool or strategy for improvement.

8.6.3 Limitations

This endeavour was limited in several ways, which are important to weigh when conducting any form of linguistic analyses. First, linguistic analyses provide important insight into the microprocesses underlying language, but present logistical challenges related to scope and breadth. This is an inherent consideration when navigating large amounts of language across contexts. Because this study focused on one case only, it is difficult to make generalizations about the use of rhetoric more broadly. By narrowing the scope to participation in particular meetings, the study did not explore more informal forms of interaction that might have yielded different insights into the principal’s use of persuasion. As such, this study and other studies are often limited by issues of accessibility and feasibility.

Second, methodologically, the study did not take a systematic approach to exploring the co-construction of meaning through argument and counter-argument that occurs through interaction. Understanding leadership as a distributed process across actors (Spillane, 2012) raises concerns about the approach that focused narrowly on an individual’s language use, with limited consideration of the influence of interaction. Exploring the possibilities of other forms of discourse analysis that take

interaction into account might provide a different form of insight into the negotiated enactment of school improvement among leaders, staff, and others. While rhetorical analyses can provide important insights into the role of persuasion, conversation analyses might help unpack the role of interaction and discussion in creating new meanings, fostering collaboration, and building consensus for improvement efforts.

Third, the rhetorical analyses conducted here drew on informal and unplanned interactions occurring within meetings. Although the meetings provided a particular, formal context for interaction, the analyzed utterances were not necessarily premeditated. Thus, researchers recognized the implicit and likely unplanned nature of leadership language here, limiting conclusions about the intentionality of the principal's use of rhetoric. This is an inherent feature of studying language in everyday practice, as opposed to more formal and prepared speech acts, such as presentations and written communications (Heracleous & Barrett, 2001). Although I have framed an argument here for the importance of examining both formal and informal linguistic structures, we need to interpret findings as they relate to the nature of the language analyzed.

Keeping these limitations in mind, I would argue that the approach outlined in detail above provides a useful model for how one might uncover, learn from, and shape the underlying rhetorical forms at play in the context of school improvement. Such analyses allow us to explore the often invisible mechanisms of language that influence the day-to-day realities of social organizations. In particular, they shine a light on the role of persuasion in leadership practice and present an opportunity for further research that builds on an understanding of how rhetorical form and content might be used to promote and develop school improvement.

8.7 Methodological Considerations

As the example discussed above demonstrates, linguistic analyses provide substantial opportunities for learning about leadership language in the context of school improvement. Even so, there are some important considerations worth exploring when thinking about these opportunities. The examples from our work draw on analyses of transcripts generated from recordings of interactions in meetings focused on individual school leaders, must be interpreted through a set of limitations that likely impacts most studies taking a similar approach. As with all research methodologies, discourse analyses applied to leadership are bounded by some practical considerations, which influence the feasibility of the work.

For example, issues of access are not inconsequential to the study of leadership language, particularly given that some of the most important moments of leadership practice occur through one-on-one interactions with staff, students, and families. These interactions are often sensitive in nature and extremely private. Researchers are unlikely to gain access to these one-on-one interactions, let alone have opportunities to digitally record such meetings for detailed analysis. As such, research on leadership language runs the risk of focusing on a narrow slice of language that is

more easily obtained, such as public communications and formal meetings. I do not intend to negate the value of linguistic analyses of these practices, but rather highlight the challenges of collecting the full repertoire of interactions relevant to understanding how leaders use language to influence practice and work toward school improvement.

Furthermore, as discussed above, it is often unfeasible to conduct large-scale studies of microprocesses of interactions. This limits the possibilities for generalizability and runs the risk of leading to a series of disjointed studies, which cannot provide wide-ranging applicability to leadership across distinct contexts. The potential to batch process larger sets of text segments or utterances continues to expand as new software technologies emerge. Even so, the sheer volume of language in practice requires carefully constructed samples focused on crafting a meaningful sample across leaders. Again, I want to be clear that there is great value to in-depth analyses of individual cases, which can illuminate undergirding structures of language use within particular contexts. I raise this consideration in order to emphasize the importance of both case selection and collaboration across researchers to compile comparable data and facilitate cross-case analyses at a larger scale.

Mixed-methods approaches also offer great potential for leveraging linguistic analyses for learning about leadership. School improvement efforts rely on complex processes occurring across organizations, and understanding them requires more than one approach to research. Often, researchers rely on survey or interview methods to provide insight into how stakeholders perceive reforms. It is more difficult to document changes to practice itself, but building on ethnographic observation, logs, and other forms of documentation have been used to that end. As discussed here, linguistic analyses offer one way to understand the mechanisms by which these changes to practice occur and therefore provide insight into how leaders actually enact shifts in both practice and perceptions. Mixed-methods approaches to studying leadership have become more widespread, as researchers bring together quantitative approaches to provide breadth with more qualitative methods to ensure depth (Tashakkori & Teddlie, 2010). Often, however, even these efforts to provide a more holistic understanding of improvement fail to account directly for the role of language, viewing language as a vehicle or medium for practice rather than an aspect of practice itself. By drawing on multiple methods to understand school improvement and incorporating rhetorical analyses, researchers will be able to better understand the relations between leadership language, educators' perspectives, and actual shifts in practice.

Considerations of feasibility, access, and generalizability are all important to future researchers committed to a linguistic turn in the study of school leadership and effectiveness. Building on a growing body of research across the fields of organization studies and education, future scholarship might leverage new analytic tools alongside longstanding linguistic methods to unpack the various ways in which language, in both formal and informal interactions, shapes the daily practices of school leaders and their staff. Through an expanding set of such studies, a collaborative, meta-analytic approach might generate opportunities for sharing across studies and the development of insight across leadership contexts and linguistic practices.

8.8 Implications for Practice

As shown above, various forms of linguistic analyses, such as rhetorical analyses, can be used to help researchers develop an understanding of how language informs, shapes and creates daily practices within schools. But the value of employing such methodologies does not end with researchers. By turning the lens on the everyday interactions that comprise our social organizations, we uncover the often invisible ways work gets done. This is important because, “the routines we practice most, and the interactions we repeatedly engage in are so familiar that we no longer pay attention to them” (Copland & Creese, 2015, p. 13). School leaders themselves have much to gain from examining their own language use and considering the implicit forms of their language within their schools and communities.

Given the context of reform in the United States, where I work, the skills of rhetoric have become all the more important to school leaders in recent years. With high-stakes accountability systems impacting schools and systems of schools, leaders play an increasingly important role in competing for resources, marketing their schools, and navigating the various conflicts that arise in a high-pressure environment (Lowenhaupt, 2014). At the same time, they are responsible for establishing a vision anchored in the professional ethos of the educational field and ensuring that they provide safe, nurturing spaces for students to inhabit (Frick, 2011). As illustrated above, leadership language has the potential to bridge these enduring norms and commitments of educators with new innovations and practices associated with school improvement. However, this is complex work, and as Gronn (1983) reminds us, talk is the work in which leaders need to engage.

Yet, as I have learned from engaging in fieldwork and working directly with current school leaders, many educational leaders do not apply a purposeful and strategic approach to much of their communication. In feedback they offer teachers, in the management of various meetings, and in day-to-day encounters in the hallway, leaders often focus on the content, rather than on the delivery of their messages. Leadership training programs and professional development opportunities might develop explicit opportunities to learn about linguistic concepts, forms of rhetoric, and a strategy for language use as it relates to supporting school improvement. By considering language as an explicit and core aspect of practice, aspiring and practicing school leaders will have an opportunity to shift their understanding towards incorporating a more purposeful approach to language use in their daily practice.

Throughout this chapter, I have sought to establish the need to leverage research methodologies that facilitate the examination of linguistic features of everyday leadership practices. Although language is a central aspect of leadership, it is often overlooked as simply the implicit medium for action. I have argued here that language use is in fact an explicit and crucial action in and of itself, and one deserving more careful attention, both as a focus for researchers and as an area of development for aspiring and practicing leaders.

References

- Alvesson, M., & Kärreman, D. (2000). Taking the linguistic turn in organizational research: Challenges, responses, consequences. *Journal of Applied Behavior Science*, 36(2), 136–158.
- Aristotle. (1992). *The art of rhetoric*. London, UK: Penguin Classics.
- Bolman, L. G., & Deal, T. E. (2003). *Reframing organizations: Artistry, choice, and leadership*. San Francisco, CA: Jossey-Bass.
- Brown, A. D., Ainsworth, S., & Grant, D. (2012). The rhetoric of institutional change. *Organization Studies*, 33(3), 297–321.
- Browne-Ferrigno, T. (2003). Becoming a principal: Role conception, initial socialization, role-identity transformation, purposeful engagement. *Educational Administration Quarterly*, 39(4), 468–503.
- Camburn, E. M., Spillane, J. P., & Sebastian, J. (2010). Assessing the utility of a daily log for measuring principal leadership practice. *Educational Administration Quarterly*, 46(5), 707–737.
- Copland, F., & Creese, A. (2015). *Linguistic ethnography: Collecting, analysing and presenting data*. Beverly Hills, CA: Sage.
- Corbett, E. P. J., & Connors, R. J. (1999). *Classical rhetoric for the modern student* (4th ed.). New York, NY: Oxford University Press.
- Darling-Hammond, L., LaPointe, M., Meyerson, D., & Orr, M. T. (2007). *Preparing school leaders for a changing world: Lessons from exemplary leadership development programs. School leadership study. Executive summary*. Palo Alto, CA: Stanford Educational Leadership Institute.
- Deal, T. E., & Celotti, L. D. (1980). How much influence do (and can) educational administrators have on classrooms? *The Phi Delta Kappan*, 61(7), 471–473.
- Deal, T. E., & Peterson, K. D. (1999). *Shaping school culture: The heart of leadership*. San Francisco, CA: Jossey-Bass.
- Dillard, C. B. (1995). Leading with her life: An African American feminist (re) interpretation of leadership for an urban high school principal. *Educational Administration Quarterly*, 31(4), 539–563.
- Firestone, W. A. (1985). The study of loose coupling: Problems, progress, and prospects. In A. Kerckhoff (Ed.), *Research in sociology of education and socialization* (Vol. 5, pp. 3–30). Greenwich, CT: JAI Press.
- Firestone, W. A., & Wilson, B. L. (1985). Using bureaucratic and cultural linkages to improve instruction: The principal's contribution. *Educational Administration Quarterly*, 21(2), 7–30.
- Frick, W. C. (2011). Practicing a professional ethic: Leading for students' best interests. *American Journal of Education*, 117(4), 527–562.
- Gee, J. P. (1999). *An introduction to discourse analysis: Theory and method*. London, UK: Routledge.
- Gill, A. M., & Whedbee, K. (1997). Rhetoric. In T. A. van Dijk (Ed.), *Discourse studies: A multi-disciplinary introduction* (Vol. 1, pp. 157–183). Beverly Hills, CA: Sage.
- Goffman, E. (1981). *Forms of talk*. Philadelphia, PA: University of Pennsylvania Press.
- Goldring, E., Huff, J., May, H., & Camburn, E. (2008). School context and individual characteristics: What influences principal practice? *Journal of Educational Administration*, 46(3), 332–352.
- Grissom, J. A., & Loeb, S. (2011). Triangulating principal effectiveness: How perspectives of parents, teachers, and assistant principals identify the central importance of managerial skills. *American Educational Research Journal*, 48(5), 1091–1123.
- Gronn, P. (1983). Talk as the work: The accomplishment of school administration. *Administrative Science Quarterly*, 28(1), 1–21.
- Hallett, T., Harger, B., & Eder, D. (2009). Gossip at work: Unsanctioned evaluative talk in formal school meetings. *Journal of Contemporary Ethnography*, 38(5), 584–618.

- Hallinger, P. (2005). Instructional leadership and the school principal: A passing fancy that refuses to fade away. *Leadership and Policy in Schools*, 4(3), 221–239.
- Heracleous, L., & Barrett, M. (2001). Organizational change as discourse: Communicative actions and deep structures in the context of information technology implementation. *The Academy of Management Journal*, 44(4), 755–778.
- Hornig, E. L., Klasik, D., & Loeb, S. (2010). Principal's time use and school effectiveness. *American Journal of Education*, 116(4), 491–523.
- Khalifa, M. (2012). A re-new-ed paradigm in successful Urban School leadership principal as community leader. *Educational Administration Quarterly*, 48(3), 424–467.
- Klar, H. W., & Brewer, C. A. (2013). Successful leadership in high-needs schools: An examination of core leadership practices enacted in challenging contexts. *Educational Administration Quarterly*, 49(5), 768–808.
- Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly*, 52(5), 711–753.
- Leithwood, K., Harris, A., & Hopkins, D. (2008). Seven strong claims about successful school leadership. *School leadership and management*, 28(1), 27–42.
- Lortie, D. C. (2009). *School principal: Managing in public*. Chicago: University of Chicago Press.
- Lowenhaupt, R. (2014). The language of leadership: Principal rhetoric in everyday practice. *Journal of Educational Administration*, 52(4), 446–468.
- Lowenhaupt, R., & McNeill, K. L. (2019). Making the case for K8 science supervision: Subject-specific instructional leadership in an era of reform. *Leadership and Policy in Schools*, 18(3), 460–484. <https://doi.org/10.1080/15700763.2018.1453937>
- Lowenhaupt, R., Spillane, J., & Hallett, T. (2016). Accountability talk: Pulling down institutional logics in organizational practice. *Journal of School Leadership*, 26(5), 783–810.
- Mehan, H. (1983). The role of language and the language of role in institutional decision making. *Language in Society*, 12(2), 187–211.
- Mouton, N., Just, S. N., & Gabrielsen, J. (2012). Creating organizational cultures: Re-conceptualizing the relations between rhetorical strategies and material practices. *Journal of Organizational Change Management*, 25(2), 315–331.
- Peterson, K. D. (1977). The Principal's tasks. *Administrator's Notebook*, 26(8), 1–4.
- Peterson, K. D., & Kelley, C. (2002). Principal in-service programs. In M. S. Tucker & J. B. Coddling (Eds.), *The principal challenge: Leading and managing schools in an era of accountability* (pp. 313–333). San Francisco, CA: Jossey-Bass.
- Rallis, S. F., & Goldring, E. B. (2000). *Principals of dynamic schools: Taking charge of change*. Thousand Oaks, CA: Corwin Press.
- Riehl, C. (1998). We gather together: Work, discourse, and constitutive social action in elementary school faculty meetings. *Educational Administration Quarterly*, 34(1), 91–125.
- Riehl, C. J. (2000). The Principal's role in creating inclusive schools for diverse students: A review of normative, empirical, and critical literature on the practice of educational administration. *Review of Educational Research*, 70(1), 55–81.
- Spillane, J., & Lowenhaupt, R. (2019). *Navigating the Principalship: Key insights for new and aspiring leaders*. Alexandria, VA: ASCD.
- Spillane, J. P. (2012). *Distributed leadership* (Vol. 4). San Francisco, CA: Wiley.
- Spillane, J. P., & Lee, L. C. (2014). Novice school principals' sense of ultimate responsibility: Problems of practice in transitioning to the principal's office. *Educational Administration Quarterly*, 50(3), 431–465.
- Stone, D. A. (1997). *Policy paradox: The art of political decision making*. New York, NY: WW Norton.
- Suddaby, R., & Greenwood, R. (2005). Rhetorical strategies of legitimacy. *Administrative Science Quarterly*, 50(1), 35–67.

- Symon, G. (2005). Exploring resistance from a rhetorical perspective. *Organization Studies*, 26(11), 1641–1663.
- Tashakkori, A., & Teddlie, C. (Eds.). (2010). *Sage handbook of mixed methods in social & behavioral research*. Thousand Oaks, CA: Sage.
- Watson, T. J. (1995). Rhetoric, discourse and argument in organizational Sensemaking: A reflexive tale. *Organization Studies*, 16(5), 805–821.
- Wolcott, H. (1973). *The man in the principal's office: An ethnography*. New York, NY: Holt, Rinehart and Winston.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 9

Designing and Piloting a Leadership Daily Practice Log: Using Logs to Study the Practice of Leadership



James P. Spillane and Anita Zuberi

9.1 Introduction

An extensive research base suggests that school leadership can influence those in-school conditions that enable instructional improvement (Bossert, Dwyer, Rowan, & Lee, 1982; Hallinger & Murphy, 1985; Leithwood & Montgomery, 1982; Louis, Marks, & Kruse, 1996; McLaughlin & Talbert, 2006; Rosenholtz, 1989) and indirectly affect student achievement (Hallinger & Heck, 1996; Leithwood, Seashore-Louis, Anderson, & Wahlstrom, 2004). Equally striking, philanthropic and government agencies are increasingly investing considerable resources on developing school leadership, typically (though not always) equated with the school principal. Taken together, these developments suggest that the quantitative measurement of school leadership merits the attention of scholars in education and program evaluation.

Rising to this research challenge requires attention to at least two issues. First, scholars of leadership and management have recognized for several decades that an exclusive focus on positional leaders fails to capture these phenomena in organizations (Barnard, 1938; Cyert & March, 1963; Katz & Kahn, 1966). Although in no way undermining the role of the school principal, this recognition argues for thinking about leadership as something that potentially extends beyond those with

This is a reprint of the article published in 2009 in *Educational Administration Quarterly*, 45(3), 375–423.

J. P. Spillane (✉)
Northwestern University, Evanston, IL, USA
e-mail: j-spillane@northwestern.edu

A. Zuberi
Duquesne University, Pittsburgh, PA, USA
e-mail: zuberia@duq.edu

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_9

155

formally designated leadership and management positions (Heller & Firestone, 1995; Ogawa & Bossert, 1995; Pitner, 1988; Spillane, 2006). Recent empirical work underscores the need for moving beyond an exclusive focus on the school principal in studies of school leadership and management and for identifying others who play key roles in this work (Camburn, Rowan, & Taylor, 2003; Spillane, Camburn, & Pareja, 2007). Second, some scholars have called for attention to the practice of leadership and management in organizations—specifically, its being distinct from an exclusive focus on structures, roles, and styles (Eccles & Nohria, 1992; Gronn, 2003; Heifetz, 1994; Spillane, 2006; Spillane, Halverson, & Diamond, 2001). The study of work practice in organizations is rather thin, in part because getting at practice is rather difficult, whether qualitatively or quantitatively. According to sociologist David Wellman, how people work is one of the best kept secrets in America (as cited in Suchman, 1995). A practice or “action perspective sees the reality of management as a matter of actions” (Eccles & Nohria, 1992, p. 13) and so encourages an approach to studying leadership and management that focuses on action rather than leadership structures, states, and designs. Focusing on leadership and management as activity allows for people in various positions in an organization to have responsibility for leadership work (Heifetz, 1994). In-depth analysis of leadership practice is rare but essential if we are to make progress in understanding school leadership (Heck & Hallinger, 1999).

This article is premised on the assumption that examining the day-to-day practice of leadership is an important line of inquiry in the field of organizational leadership and management. One key challenge in pursuing this line of inquiry involves the development of research instruments for studying the practice of leadership in large samples of schools. This article reports on one such effort—the design and piloting of a Leadership Daily Practice (LDP) log—which attempts to capture the practice of leadership in schools, with an emphasis on leadership for mathematics instruction in particular and leadership for instruction in general. Based on a distributed perspective (Spillane et al., 2007), our efforts move beyond an exclusive focus on the school principal, in an effort to develop a log that generates empirical data about the interactions of leaders, formal and informal, and their colleagues.

Our article is organized as follows: We begin by situating our work conceptually and methodologically and by examining the challenges of studying the practice of leadership. Next, we consider the use of logs and diaries to collect data on practice, and we describe the design of the LDP log. We then describe our method. Next, we organize our findings based on the validity of the inferences that we can make given the data generated by the LDP log—specifically, around four research questions:

- *Question 1:* To what extent do study participants consider the interactions that they enter into their LDP logs to be *leadership*, defined as a social influence interaction?
- *Question 2:* To what extent are study participants’ understandings of the constructs (as used in the log to describe social interactions) aligned with researchers’ definitions of these constructs (as defined in the log manual)?

- *Question 3:* To what extent do study participants and the researchers who shadowed them agree when using the LDP log to describe the same social interaction?
- *Question 4:* How representative are study participants' log entries regarding the types of social influence interactions recorded by researchers for the same logging days?

Research Questions 1 and 2 can be thought of in terms of construct validity for two reasons: First, we examine whether interactions selected by study participants for inclusion in the log are consistent with the researchers' definition and operationalization of leadership as a social influence interaction (as denoted in the LDP log and its accompanying manual). Second, we examine the extent to which study participants' understandings of key terms (as used in the log to describe these interactions) align with researchers' definitions (as outlined in the log manual). Research Question 3 examines the magnitude of agreement between the log entries of the study participants and the entries of the observers who shadowed them regarding the same social influence interaction. We can think about this interrater reliability between loggers and researchers for the same interaction as a sort of concurrent validity; that is, it focuses on the agreement between two accounts of the same leadership interaction. Research Question 4 centers on a threat to validity, introduced because study participants selected one interaction per hour for entry into their LDP logs (rather than every interaction for that hour); hence, we worry that study participants might be more prone to selecting some types of social influence interactions over others. To examine the threat of selection bias, we investigate whether the interactions that study participants logged were representative of all the interactions they engaged in, as documented by researchers who recorded every social interaction on the days that they shadowed select participants. We conclude with a discussion of the results and with suggestions for redesigning the LDP log. We should note that our primary concern in this article is the design and piloting of the LDP log. Thus, we report here the substantive findings only in the service of discussing the validity of the LDP log, leaving for another article a comprehensive report on these results.

9.2 Situating the Work: Conceptual and Methodological Anchors

9.2.1 Conceptual Anchors

We use a distributed perspective to frame our investigation of school leadership (Gronn, 2000; Spillane, 2006; Spillane et al., 2001). The distributed perspective involves two aspects: the leader-plus aspect and the practice aspect. The leader-plus aspect recognizes that the work of leadership in schools can involve multiple people. Specifically, people in formally designated leadership positions and those

without such designations can take responsibility for leadership work (Camburn et al., 2003; Heller & Firestone, 1995; Spillane, 2006).

A distributed perspective also foregrounds the practice of leadership; it frames such practice as taking shape in the interactions of leaders and followers, as mediated by aspects of their situation (Gronn, 2002; Spillane, Halverson, & Diamond, 2004). Hence, we do not equate leadership practice with the actions of individual leaders; rather, we frame it as unfolding in the interactions among school staff. Efforts to understand the practice of leading must pay attention to interactions, not simply individual actions. Foregrounding practice is important because practice is where the rubber meets the road—"the strength of leadership as an influencing relation rests upon its effectiveness as activity" (Tucker, 1981, p. 25).

Similar to others, we define leadership as a social influence relationship—or perhaps more correctly (given our focus on practice), an influence interaction (Bass, 1990; Hollander & Julian, 1969; Tannenbaum, Weschler, & Massarik, 1961; Tucker, 1981). We define leadership practice as those activities that are either understood by or designed by organizational members to influence the motivation, knowledge, and practice of other organizational members in an effort to change the organization's core work, by which we mean teaching and learning—that is, instruction.

9.2.2 Methodological Anchors

With a few exceptions (e.g., Scott, Ahadi, & Krug, 1990), scholars have relied mostly on ethnographic and structured observational methods (e.g., shadowing) or annual questionnaires to study school leadership practice (Mintzberg, 1973; Peterson, 1977). Although both approaches have strengths, they have their limitations. Similar to ethnography, structured observations have the benefit of being close to practice. Unlike ethnography, this approach hones in on specific features of practice and the environment, thereby resulting in more focused data (Mintzberg, 1973; Peterson, 1977). Ethnography and structured observations (although close to practice) are costly, and such large-scale studies are typically too expensive to carry out in more than a few schools, especially under the presumption that leadership extends beyond the work of the person in the principal's office.

Surveys are a less expensive option than structured or semistructured observations; they are cheap to administer, and they generate data on large samples. However, some scholars question the accuracy of survey data with respect to practice, as being distinct from attitudes and values. Specifically, recall of past behavioral events on surveys can be difficult and can thus lead to inaccuracies (Tourangeau, Rips, & Rasinski, 2000). Inaccuracy is heightened as time lapses between the behavior and the recording of it (Hilton, 1989; Lemmens, Knibbe, & Tan, 1988; Lemmens, Tan, & Knibbe, 1992).

Diaries of various sorts offer yet another methodological approach for studying leadership practice, including event diaries, daily logs, and Experience Sampling Method (ESM) logs. Event diaries require practitioners to record when an event

under study happens (e.g., having a cigarette). Daily logs require practitioners to record, at the end of the day, the events that occurred throughout the day. ESM logs beep study participants at random intervals during the day, cueing them to complete a brief questionnaire about what they are currently doing. Among the advantages of the ESM methodology is that (a) practitioners can report on events when they are fresh in their minds, (b) they do not have to record every event, and (c) the random design allows for a generalizable sample of events (Scott et al., 1990). The ESM methodology, however, is intrusive, and participants can be beeped while engaged in sensitive matters.

The evidence suggests that logs provide a more accurate measure of practice than that of annual surveys, although most of this work has not centered on leadership practice (Camburn & Han, 2005; Mullens & Gaylor, 1999; Smithson & Porter, 1994). The work reported here builds on the log methodology by describing the design and pilot study of the LDP log in particular.

9.3 Designing the LDP Log

Our development of the LDP log was prompted by earlier work on the design of an End of Day log and an ESM log, both of which focused on the school principal's practice (Camburn, Spillane, & Sebastian, 2006). The ESM log informed our design of the LDP log; so, we begin with a description of that process and then turn to the LDP log design.

9.3.1 *ESM Log Design*

A prototype of the ESM log was based on a review of the literature on the ESM approach and school leadership. Developed with closed-ended items, the ESM log probed several dimensions of practice, including the focus of the work, where it happened, who was present, and how much time was involved. Open-ended log items place considerable response burden on participants who have to write out responses; they also pose major challenges for making comparisons across participants (Stone, Kessler, & Haythornthwaite, 1991). Hence, in designing the ESM log, we created closed-ended items (given on our review of the literature) and then refined them in three ways. First, we used the items to code ethnographic field notes on school administrators' work, exploring the extent to which our items captured what was being described in the notes. Second, we had 11 school leadership scholars critique the items.

After performing these two steps, we revised our items and subsequently conducted a preliminary pilot of the EMS log with five Chicago school principals over 2 days. Each principal was shadowed under a structured protocol over the 2-day period as they completed the ESM log when beeped at random intervals. We again

revised the log on the basis of an analysis of these data; as a result, we added a series of affect questions to tap participants' moods. In spring 2005, we conducted a validity study of the ESM log with 42 school principals in a midsize urban school district. Overall, this work suggested that the log generated valid and reliable measures on those dimensions of school principal practice that it measured.

9.3.2 LDP Log Design

The ESM log had some limitations, which prompted our efforts to design a LDP log. To begin with, we wanted to move beyond a focus on the school principal, to examine the practice of other school leaders. Data generated by the ESM log on 42 school principals showed that others—some with formally designated leadership positions and others without (and often with full-time teaching responsibilities)—were important to understanding leadership, even when measured from the perspective of the school principal's workday. Using the ESM log with those who were teaching most or all of the time posed a challenge, owing to the random-beeping requirement. Furthermore, we wanted to zero in on leadership interactions, but the ESM log did not enable us to distinguish leadership interactions from management or maintenance interactions. Hence, we designed the LDP log to be used with a wider spectrum of leaders (including those with full-time teaching responsibilities) and to focus on leadership (defined as social influence interactions).

At the outset, we developed a prototype of the LDP log, based on the ESM log and with input from scholars of teaching and school leadership. Using this prototype, we then conducted a focus group with teams of school leaders from three schools, which raised several issues that subsequently informed the redesign of the LDP log. First, participants in the focus group thought that a randomly beeping paging device (to remind them to log an interaction) would be too intrusive. Moreover, we were not convinced that random beeping would enable us to capture leadership interactions (especially for school staff with full-time classroom teaching responsibilities), namely, because these events might be rare; as such, there would be little chance that the signal and the event would coincide (Bolger, Davis, & Rafaeli, 2003; Wheeler & Reis, 1991). Furthermore, leadership interactions were likely to be unevenly distributed across the day (especially for those who taught full-time)—that is, occurring between classes or at the end or beginning of the school day.

Focus group participants also suggested that it would be too onerous to record all interactions related to leadership (i.e., for mathematics in particular and for classroom instruction in general). Hence, to reduce the reporting burden on study participants, we decided that they would select only one interaction (of potentially numerous interactions) from each hour between 7 a.m. and 5 p.m. and report on these selected interactions on a Web-based log at the end of the workday. When multiple interactions occurred in an hour, respondents were instructed to choose the interaction that was most closely related to mathematics instruction and, if nothing was related to mathematics, an interaction most closely tied to curriculum and

instruction. Although we acknowledge that the work of school staff is not limited to the official school day, we decided that adding at least 1 h before and after the school day would capture some of the interactions that take place during such time, without burdening respondents at home. Standardizing hours in this way facilitates comparisons across respondents and schools because all study participants are asked to report on the same periods. We acknowledge the limitations of this approach in terms of a qualitative or interpretive perspective.

The decision to have study participants complete the LDP log at the end of the day posed a second design challenge in that we needed to minimize recall bias, which might have been introduced from having study participants make their log entries several hours after the occurrence of the interaction (Csikszentmihalyi & Larson, 1987; Gorin & Stone, 2001). Earlier work comparing data based on the ESM log (in which participants made entries when beeped) to data generated by an End of Day log (where participants made entries at the end of the day) suggested high agreement between the two data sources on how school principals spent their time (Camburn et al., 2006). The LDP log, however, probed several other dimensions of practice, including who was involved and what the substance of the interaction was. To minimize recall bias, we create a paper log that participants could use to track their interactions across the workday. Focus group participants were split on the design of these logs, with some preferring checklists and with others arguing for blank tables for jotting reminders. We designed the paper log so that participants could choose one of these options.

In another design decision, we opted for mostly closed-ended questions, with a few open-ended ones. We used many of the ESM items as our starting point for generating the stems for the closed-ended items (see Appendix A). Three additional issues informed the design of the log. First, we asked respondents to report if the day was typical. Second, we asked respondents if they used the paper log to record interactions throughout the day. Third, we asked respondents to identify whether the interaction being logged was intended to influence their knowledge, practice, and motivation. To help minimize differences in interpretation, we worked with study participants on the meaning of each concept and provided them with a manual to help them to decide whether something was about knowledge, practice, or motivation.¹ To help maintain consistency across respondents, the manual defined an *interaction* as “each new encounter with a person, group, or resource that occurs in an effort to influence knowledge, practice, and motivation related to mathematics or curriculum and instruction.” To simplify our pilot study, we asked study participants not to report on interactions with students and parents.

Loggers were asked at the outset if the interaction involved an attempt on their part to influence someone (i.e., provide) or an attempt to be influenced (i.e., solicit;

¹ The Leadership Daily Practice (LDP) log states that knowledge refers to “interactions re-garding information, what you learned, and specific content”; practice includes “what you do, daily activities, teaching, and pedagogy”; and motivation refers to “support, encouragement, and the provision of resources.” The instruction manual for the LDP log also provides some examples of how to use these categories.

see Appendix A).² Depending on whether respondents selected *provide* or *solicit*, they followed one of two paths through the log. Questions were similar but tailored to whether the respondent was in the role of leader or follower in the interaction. We also designed the LDP log to capture whether an interaction was planned or spontaneous. Prior research suggests that many of the interactions in which school leaders engage are spontaneous (Gronn, 2003). To help respondents decide whether an interaction was planned or spontaneous, respondents were told to evaluate whether the following criteria were predetermined: participants, time, place, and topic.³ The log also asked respondents to estimate, at the end of the day, the amount of time they spent doing various tasks for that day. Tasks were split into four broad categories: administrative duties (school, department, and grade), curriculum and instructional leadership duties, classroom teaching duties, and nonteaching duties. As noted earlier, our LDP log categories were derived from earlier work on the End of Day and ESM logs, as well as from our review of the literature and from the input of scholars.

9.4 Research Methodology

We used a triangulation approach (Camburn & Barnes, 2004; Campbell & Fiske, 1959; Denzin, 1989; Mathison, 1988) to study the validity of the LDP log. Specifically, we used multiple methods and data sources (Denzin, 1978), including logs completed by study participants as well as observations and cognitive interviews conducted by researchers.

For a 10-day period during fall 2005, study participants from four urban schools were asked to log one interaction per hour that was intended to influence their knowledge, practice, or motivation or in which they intended to influence the knowledge, practice, or motivation of a colleague. Participants were also asked to note what prompted the interaction, who was involved, how it took place, what transpired, and what subject it pertained to (see Appendix A). Two schools were middle schools (Grades 6–8) and two were combined (Grades K–8).

9.4.1 Sample

Sampling leaders is complex when based on a distributed perspective on school leadership. To begin with, we selected all the formally designated leaders who might work on instruction, including principals, assistant principals, and curriculum

²In cases where several topics may be discussed in one interaction, participants are asked to “please consider who initiated interaction.”

³The log offers the following instructions: “In order to determine if an interaction was planned or spontaneous, please consider if the participants, time, place and topic were pre-determined before the interaction took place. If all four conditions apply, code the interaction as planned.”

specialists for mathematics and literacy. We also wanted to sample informal leaders, those identified by their colleagues as leaders but who did not have formally designated leadership positions. To select informal leaders, we used a social network survey, designed to identify school leaders. Specifically, informal leaders were defined as those teachers who had high “indegree” centrality measures, based on a network survey administered to all school staff. *Indegree centrality* is a measure of the number of people who seek advice, guidance, or support from a particular actor in the school. Hence, school staff with no formal leadership designation but with high indegree centrality scores also logged and were thus shadowed in our study. Furthermore, we asked all the mathematics teachers to log (regardless of indegree centrality).

One-on-one or group training was provided to familiarize participants with the questions on the log and the definitions of key terms. Each participant was then provided with the LDP log’s user manual. All together, 34 school leaders and teachers were asked to complete the LDP log to capture the nature of their interactions pertaining to leadership for curriculum and instruction over a 2-week period (specifically, 4 principals, 4 assistant principals, 1 dean of students, 3 math specialists, 4 literacy specialists, and 18 teachers). The overall completion rate showed that, on average, participants completed the log for 68% of the days (i.e., 6.8 out of 10 days; see Table 9.1). This figure varied substantially by role, from a low of 30% (for principals) to a high of 95% (for literacy specialists).⁴ Whereas the overall response rate is good, the response rate for principals is low. Although there was some variation among principals, the range was from 0% to 70%. The average number of interactions that individuals logged per day (only counting those who completed the log for the day) declines over the 2-week period (see Fig. 9.1), ranging from a high of 3.0 (on the first Tuesday of logging) to a low of 1.4 (on the last logging day, the second Friday). Of the 34 study participants, 22 were shadowed across all four schools over the 2-week logging period. The group who was shadowed consisted of all the principals ($n = 4$), math specialists ($n = 3$), and literacy specialists ($n = 4$) in the logging sample, as well as all but one of the assistant principals ($n = 4$). Only teacher leaders ($n = 7$) were shadowed; as such, the response rate of this group was 74%, slightly higher than the 66% for all the teachers who completed the LDP log (see Table 9.2). Shadowing may have increased the likelihood of log completion among this group, but our data do not permit an investigation into the issue.

Compared to all loggers, the shadowed respondents logged slightly more interactions on average per day (see Fig. 9.2). This is not surprising, given that we purposefully shadowed the formal and informal leaders in the schools, whom we expected to have more interactions to report. The shadowing process, as followed by the cognitive interviews, may have also contributed to the higher number of interactions logged by these participants. As with the full sample, the average number of

⁴Numerous participants stated that they did not complete the log in the evening, because they were preoccupied watching the baseball game (i.e., data collection occurred during the World Series).

Table 9.1 Response rates for leadership daily practice log

	Participants	Potential days Logged	Actual days logged	% of potential days logged
	<i>n</i>	<i>n</i>	<i>n</i>	%
Total	34	340	230	67.6
School				
Acorn	7	70	54	77.1
Alder	10	100	67	67.0
Ash	10	100	58	58.0
Aspen	7	70	51	72.9
Role				
Principals	4	40	12	30.0
Assistant principals	5	50	36	72.0
Mathematics specialists	3	30	26	86.7
Literature specialists	4	40	38	95.0
Teachers	18	180	118	65.6

interactions reported each day peaked early in the first week and dipped by the end of the second week.

Nineteen study participants were shadowed for 2 days each, whereas three participants were shadowed for only 1 day. We have log entries for 30 of 41 days during which study participants were shadowed. Only three of the shadowed study participants were missing entries for all the days during which they were shadowed (one principal, one assistant principal, and one teacher). Our analysis is therefore based on the shadow data and log entries for 19 people across four schools. The response rate for completing the LDP log when being shadowed was 73%, which is slightly higher than that of the entire logging period (see Table 9.3).

9.4.2 Data Collection

Observers who shadowed study participants recorded observations throughout the day on a standardized chart (see Appendix B). Observers were instructed to record all interactions throughout the day, with *interaction* defined as any contact with another person or inanimate object. Observers recorded interactions on a form with prespecified categories for recording (per interaction) what happened, where it took place, who it was with, how it occurred, and the time. “What happened” consisted of a substantive and subject-driven description of the interaction. Observers also recorded activity type, whether it was planned or spontaneous, and whether the observed person was providing or soliciting information. In addition, observers were beeped every 10 min to record a general description of what was going on at the time.

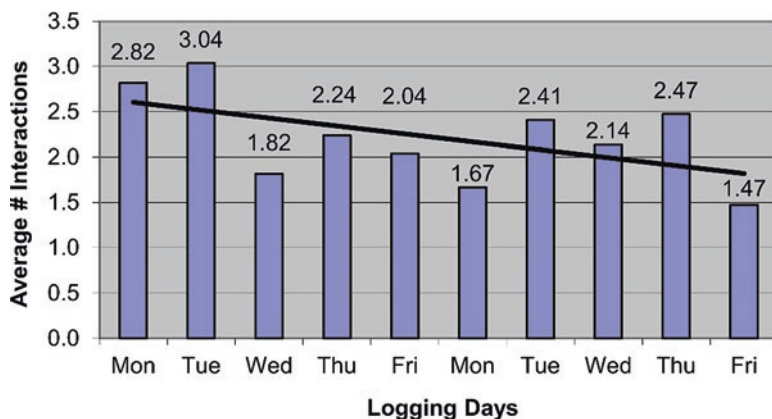


Fig. 9.1 Average interactions per day

Table 9.2 Log response rates for shadowed group (during all log days)

	People shadowed	Potential days logged	Actual days logged ^a	% of potential days logged
	<i>n</i>	<i>n</i>	<i>n</i>	%
Total	22	220	155	70.5
School				
Acorn	5	50	41	82.0
Alder	5	50	38	76.0
Ash	7	70	44	62.9
Aspen	5	50	32	64.0
Role				
Principals	4	40	12	30.0
Assistant principals	4	40	27	67.5
Mathematics specialists	3	30	26	86.7
Literature specialists	4	40	38	98.0
Teachers	7	70	52	74.3

^aShadow days only

At the end of each day of shadowing, the researcher conducted a cognitive interview with the individual being shadowed, to investigate his or her understanding of what he or she was logging and thinking about these interactions (see Appendix C). At the outset of the cognitive interview, participants were asked about their understandings of the key constructs in the LDP log. Next, they were asked to describe three interactions from that day that they recorded in the LDP log and to talk aloud about how they decided to log each interaction, focusing on such issues as whether they characterized the interaction as leadership, what the direction of influence was, and whether the interaction was spontaneous or planned. Participants were also

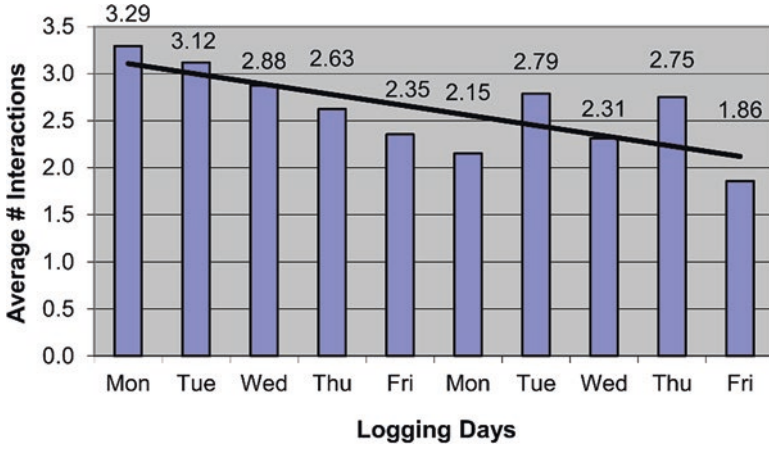


Fig. 9.2 Average interactions per day – shadowed group only

Table 9.3 Response rates for log during shadowing

	People shadowed	Potential days logged	Actual days logged ^a	% of potential days logged
	<i>n</i>	<i>n</i>	<i>n</i>	%
Total	22	41	30	73.2
School				
Acorn	5	10	8	80.0
Alder	5	10	9	90.0
Ash	7	11	7	63.6
Aspen	5	10	6	60.0
Role				
Principals	4	7	3	42.9
Assistant principals	4	8	4	50.0
Mathematics specialists	3	6	5	83.3
Literature specialists	4	8	8	100.0
Teachers	7	12	10	83.3

^aShadow days only

asked about the representativeness of their log entries. A total of 40 cognitive interviews with 21 participants were audiotaped and transcribed.

9.4.3 Data Analysis

A concern with any research instrument is the validity of the inferences that one can make based on the data that it generates about the phenomenon that it is designed to investigate. As such, our analysis was organized around four research questions that focused on whether our operationalization of leadership in the LDP log actually captured this phenomenon as we defined it (i.e., as a social influence interaction). In other words, did our attempt to operationalize and translate the construct of leadership through the questions in the LDP log work? Did the items on the LDP log capture leadership, defined as a social influence interaction?

Research Questions 1 and 2 Concerned with construct validity, we analyzed data from 40 cognitive interviews of 21 study participants, to examine their understandings of key concepts used in the LDP log to access social influence interactions (e.g., knowledge, practice) and describe or characterize such interactions (e.g., planned versus spontaneous). We also explored whether participants believed that the LDP log captured leadership, by analyzing the agreement (or lack thereof) between participants' understandings and the LDP log's user manual definition of leadership (again, as a social influence interaction).

Research Question 3 We also compared the interrater reliability between loggers and researchers for the same interactions, a form of concurrent validity. Eighty-nine entries coincided with days on which participants were shadowed, ranging from 18 to 26 log entries across schools, with a mean of 22.3 per school (see Table 9.4). Seventy-one of these entries were verifiable (i.e., the shadower recorded the interaction as well), ranging from 14 to 24 across schools, with a mean of 17.8 per school. Missing interactions from shadowers' field notes were mostly due to timing; that is, the interactions happened before school started or after it had ended, times when the shadower was not present (see Appendix D).

We examined the extent to which shadowers' data entries agreed with the data entries in the LDP log for the 71 verifiable interactions (1 = matching, 0 = non-matching), calculating the percentage of responses where the participant and the observer agreed. If there was not enough information to decide whether there was a match, then this was noted. In the case of the *what happened* category, this occurred for 7 out of 64 matches. For the *who* category, a less conservative approach was used in matching responses; namely, if one person reported the name of a teacher and the other simply reported "teacher", then this was counted as an agreement (i.e., as long

Table 9.4 Leadership daily practice log: shadow validation, sample descriptive statistics

	Total logged interactions ^a	Verifiable interactions ^b	Not able to verify
	n (%)	n (%)	n (%)
Total	89 (100.0)	71 (79.8)	18 (20.2)
School			
Acorn	26 (29.2)	24 (92.3)	2 (7.7)
Aspen	21 (23.6)	16 (76.2)	5 (23.8)
Ash	18 (20.2)	17 (94.4)	1 (5.6)
Alder	24 (27.0)	14 (58.3)	10 (41.7)
Role			
Principals	4	2 (50.0)	2 (50.0)
Assistant principals	12	11 (91.7)	1 (8.3)
Mathematics specialists	17	17 (100.0)	0 (0.0)
Literature specialists	29	24 (82.8)	5 (17.2)
Mathematics teachers	27	17 (63.0)	10 (37.0)

^aNumber of interactions logged by shadowed sample

^bRecorded in the participant's log and by the observer

as the roles matched).⁵ To account and adjust for chance agreement, we calculated the kappa coefficient where possible (i.e., for the where, how, and time of interaction), using the statistical program Stata. If a kappa coefficient is statistically significant, then “the pattern of agreement observed is greater than would be expected if the observers were guessing” (Bakeman & Gottman, 1997, p. 66). A kappa greater than .70 is a good measure of agreement; above .75 is excellent (Bakeman & Gottman, 1997; Fleiss, 1981).⁶ (See Appendix F)

Research Question 4 A key design decision with the LDP log involved having loggers select a single interaction from potentially multiple interactions per hour. Hence, a potential threat to the validity of the inferences that we can make (based on the data generated by the LDP log) is that study participants are more likely to select some types of interactions over others. As such, the LDP log data would overrepresent some types of leadership interactions and underrepresent others.

To examine how representative the interactions that study participants selected were to the population of interactions, we compared their log entries for the days on which they were shadowed to all the interactions related to mathematics and/or curriculum and instruction recorded by observers on the same days.⁷ Given that observers documented every interaction that they observed, we can regard the shadow data as an approximation for the population of interactions. Interactions were coded on

⁵ See Appendix E for a description of what constituted a match and a vague match for these codes.

⁶ Bakeman and Gottman (1997) suggest that kappas less than .70 (even when significant) should be regarded with some concern. The authors cite Fleiss (1981) who “characterizes kappas of .40 to .60 as fair, .60 to .75 as good, and over .75 as excellent” (p. 218).

⁷ The data used in this analysis are limited to days in which the study participant made at least one LDP log entry.

the basis of where, how, when, what (i.e., the subject of the interaction), and with whom. As such, we examined whether loggers were more likely to select some types of interactions over others by calculating the difference between the characteristics of logger interactions and shadower interactions and by testing for statistically significant differences.⁸

9.5 Findings

The primary goal of the work reported here involved the validity of the inferences that we can make based on the data generated by the LDP log. Specifically, we want to make inferences based on what happened to study participants, in the real world, with respect to leadership (defined as a social influence interaction). We asked participants to report on certain interactions, and the LDP log data constitute their reports of what they perceived as having happened to them. Our ability to make valid inferences from these reports depends to a great extent on how participants understood the constructs about which they were logging. If study participants understood the key constructs or terms in different ways, then we would not have comparable data across the sample, thus undermining the validity of any inferences that we might draw. As a construct, leadership is open to multiple interpretations, and it is difficult to define clearly and concretely (Bass, 1990; Lakomski, 2005). Hence, an important consideration is the correspondence between (a) study participants' understandings of the terms used to access leadership and characterize or describe it as a social influence interaction and (b) the operational definitions of these terms in the log (Research Questions 1 and 2).

Another consideration with respect to the validity of the inferences that we can make from the LDP log data concerns the extent to which the interactions logged by study participants correspond to what actually happened to them in the real world. We sought to describe what happened to study participants through field notes taken by researchers who shadowed a subsample of participants on some of the days that they completed the LDP log. Although the researchers' field notes are just another take on what happened to the study participants on the days that they were shadowed, they do represent an independent account of what the study participants did on these days (Research Question 3). Gathering comparable data with logs is challenging because study participants themselves select the interactions to log. Hence, another threat to validity involves the potential for sampling bias on the part of loggers (Research Question 4).

⁸We calculated z scores for proportions, to test whether the difference was statistically significant.

9.5.1 Research Question 1

To what extent do study participants consider the interactions that they enter into their LDP logs to be *leadership*, defined as a social influence interaction? The LDP log was designed to capture the day-to-day interactions that constitute leadership, defined as a social influence interaction. Participants reported that 89% of the interactions that they selected to log were leadership for mathematics and/or curriculum and instruction.⁹ For example, a literacy specialist confirmed that one of the interactions that he had selected involved leadership for curriculum and instruction:

I think both of us saw the need for change so we would've changed anyway but my suggestion influenced him to change the way I wanted it to. Using my background and my experience teaching literature circles I'm seeing that this isn't working certainly and giving him a different way to do it. (October 20, 2005)

Study participants overall, though critical of some of the LDP log's shortcomings, expressed satisfaction with the instrument. As one participant put it, "sometimes it's not being as accurate as I want it to be. And so probably I'd say on a 90% basis that it's accurate" (October 28, 2005). We might regard this as a form of face validity.

Part of the rationale that some study participants offered for justifying a social interaction as an example of leadership had to do with the role or position of one of the people involved. Sometimes this had to do with a formally designated position, such as a literacy specialist or a mathematics specialist. After confirming that an interaction was an example of leadership, a literacy specialist remarked,

Because the roles, although we step into different roles throughout the day, one of her roles is the curriculum coordinator and she provided materials that go with my curriculum and was able to present them to me and say, "This is done for you." My role is to then take those materials and turn it into a worthwhile lesson. So I'm not wasting my time spinning my wheels making up these game pieces; it's done. (October 26, 2005)

This participant pointed to the interaction as an example of leadership not only because it influenced his practice but because the person doing the influencing was a positional leader. The participant's remark that "although we step into different roles throughout the day" suggests that school staff can move in and out of formally designated leadership positions. A related explanation concerns the fact that a participant in an interaction was a member of a leadership team; that is, a mathematics teacher remarked, "She's part of our math leadership team too" (October 21, 2005).

Especially important from a validity perspective—given that our definition of leadership did not rely on a person's having a formally designated leadership position—participants' explanations for a leadership interaction went beyond citing formally designated positions to referring to aspects of the person who was doing the influencing. A math teacher, for example, remarked, "She influences me because I

⁹In each interview, the interviewee selected three interactions that he or she planned to enter into the log for that day, and the interviewee asked a series of structured questions about each interaction.

have respect for the person that she is and her dedication to the work that she's doing. So in that sense we work together. Because of the mutual respect and the willingness to work together, I mean there's another part of that leadership idea" (October 26, 2005). This comment suggests that the LDP log items prompt study participants to go beyond a focus on social influence interactions with those in formally designated leadership positions.

The Sampling Problem More than half the sample (56%) thought that the log accurately captured the nature of their social interactions for the day, as related to mathematics or curriculum and instruction. One mathematics teacher remarked, "The only way to better capture it is to have someone watch me or to videotape me" (October 26, 2005). Another noted, "It will probably accurately reflect the math leadership in this school... [What] it will reflect is that it's kind of happening in the halls. ...it'll probably be reflected that the majority of this is spontaneous" (October 21, 2005). These mathematics teachers' responses suggest that the LDP adequately captured the informal, spontaneous interactions that are such a critical component of leadership in schools but often go unnoticed because they are so difficult to pick up.

Still, 75% of the participants felt believed that their log entries failed to adequately portray their leadership experiences with mathematics or curriculum and instruction throughout the school year. These participants suggest two reasons why their LDP log entries did not accurately reflect their experience with leadership in their daily work—namely, because of sampling and the failure of the log to put particular interactions into context.

In sum, 9 of the 20 participants who spoke to the issue of how the log captured their leadership interactions over a school year emphasized that logging for only 2 weeks would not capture their range of leadership interactions—that is, the sampling frame of 2 consecutive weeks is problematic. Specifically, participants reported that leadership for mathematics or curriculum and instruction changes over the school year, depending on key events such as the beginning of the school year preparation, standardized testing administration, and school improvement planning.

Hence, logging for 2 weeks (10 days in total) failed to pick up on seasonal changes in leadership across the school year, and it failed to capture events that occurred monthly, quarterly, and even annually. An assistant principal explained, "I think like in the beginning, like the few weeks of school as we start to get set up for the whole school year, you know, we tend to be more busy with curriculum issues" (October 24, 2005). A mathematics specialist at a different school reported,

Well again, sometimes I'm doing much more with leadership than I have been in the last week and maybe even next week you know. When it comes time to inventory in the school, finding out curriculum, talking with different math people, consulting different books then I would have to say that at those times I'm doing more with leadership than I am in these 2 weeks here. (October 20, 2005)

Study participants pointed to specific tasks that come up at different times in the school year that were either overrepresented or not captured in the 2-week logging

period, such setting up after-school programs and organizing the science fair. The issue here concerns how we sample days for logging across the school year.

Some study participants expressed concern with respect to how interactions were sampled within days. Two participants reported that sampling a single interaction each hour was problematic. A literature specialist captured the situation:

The problem with it is sometimes there are multiple interesting experiences in a one hour time period. And so it's a definite snapshot. ...I almost wish I could choose from the entire day what was most influential so that I'm not limited by each hour what was most. (October 25, 2005)

This comment suggests that the most interesting social influence interactions may be concentrated in particular hour-long periods—many of which are not recorded, because loggers only sample a single interaction from each hour. A strategy of sampling on the day, rather than on the hour, would allow such interactions to be captured.

The concentration of social interactions at certain times of the day may be especially pronounced for formally designated leaders who teach part-time. A math specialist remarked,

I mean it might capture some of the interactions but. .. you're only allowed to insert one thing per hour. .. and I may talk to 10 people in an hour sometimes. Normally those say 3 hours that I'm teaching I don't have a lot of interaction with teachers per say unless they come in to ask me a question. It's the times that I don't [teach], you know, when I'm standing in the lunchroom and five teachers come talk to me about certain things, or I'm walking down the hall and this teacher needs this, that, and the other. (October 19, 2005)

For this specialist, social influence interactions were concentrated in her non-teaching hours, with relatively few social influence interactions during teaching hours. Hence, allowing participants to sample from the entire day, as opposed to each hour of the day, may capture more of the interactions relevant to leadership practice. For at least some school staff, key interactions may be concentrated in a few hours, such as during planning periods, and may thus be underrepresented by a sampling strategy that focused on each hour. Still, the focus on each hour may enable recall. A teacher remarked,

Well, what's nice about the interaction log is that it asks you for specific times you know the day by hours. And so it makes you really look back at your day with a fine toothcomb and say, "Okay, what exactly you know was I doing?" And then you don't realize how many interactions you really do have until you fill it out. Then you think, "Wow, I didn't think I really had that many interactions" but now that I'm filling it out I actually do interact a lot with my colleagues. (October 20, 2005)

And a literacy specialist noted, "Yeah. It's giving a good snapshot of the stuff you know or the parts of the day that I actually do work with it. ... I have to keep thinking about that time slot thing" (October 25, 2005). These comments suggest that although having participants select a single interaction for each hour has a downside, it does have an upside in that it enables their recall by getting them to systematically comb their workday.

Situating Sample Interactions in Their Intentional and Temporal Contexts Four participants suggested that the LDP log did not adequately capture leadership practice, because it failed to situate the logged interactions in their intentional and temporal contexts. An eighth-grade mathematics teacher remarked, “You need a broader picture of what I’m doing and that means the person I am and where I’m coming from as well as the goals that I have, either professionally or personally” (October 26, 2005). For this participant, the key problem was that the log failed to capture how the interactions that he logged were embedded in and motivated by his personal and professional goals and intentions. Study participants also suggested that the LDP log did not capture the ongoing nature of social influence interactions. One participant noted,

[Leadership is] gonna be ongoing. Like I was talking about with Mr. Olson, the thing we were doing today has been going on since Monday and piecing it together and looking and there’s just some other things that we have done. (October 21, 2005)

For this literature specialist, the LDP log did capture particular interactions, but it failed to allow for leadership activities that might span two or more interactions during a day or week, thereby preventing one from recording how different interactions were connected.

9.5.2 Research Question 2

To what extent are study participants’ understandings of the constructs (as used in the log to describe social interactions) aligned with researchers’ definitions of these constructs (as defined in the log manual)?

As noted above, identifying leadership as social influence interactions via the LDP log is one thing; a related but different matter lies in describing or characterizing such interactions. The validity of the inferences that we can make from the LDP log data about the types of social influence interactions in which study participants engaged depends on the correspondence between their understandings of the terms used to characterize the interactions and the operational definitions of these terms as delineated in the log manual. We designed the LDP log to characterize various aspects of social influence interactions, including the direction of influence and whether it was planned or spontaneous. If study participants’ understandings of the terminology used to operationalize these distinctions differed from one another, it would undermine the validity of the inferences that we might draw. Although our analysis suggests considerable agreement between study participants’ understandings and the definitions used in the log manual, we found that the former did not correspond to the latter for three key concepts (see Table 9.5). Specifically, participants struggled with the term *motivation*; they had difficulty deciding on the direction of influence; and they found it problematic to distinguish planned and spontaneous interactions.

Table 9.5 Cognitive interview evaluation of the leadership daily practice log

Question	Yes/ Match <i>n</i>	No/Non match <i>n</i>	Yes/ Match %
Capturing leadership			
Is this interaction an example of leadership?	89	11	89
Does the log capture the nature of your interactions for the day?	18	14	56
Does the log capture leadership throughout the school year?	7	21	25
Defining concepts			
Knowledge	19	1	95
Practice	17	3	85
Motivation	18	2	90
Describing interactions			
Did this interaction influence your knowledge?	51	7	88
Did this interaction influence your practice?	65	9	88
Did this interaction influence your motivation?	33	19	63
Did you provide information or advice?	37	9	80
Did you solicit information or advice?	35	11	76
Was this interaction planned or spontaneous?	58	38	60

Note: The totals between rows differ depending on whether the question was asked of the individual or the interaction. The totals also differ because characteristics were evaluated only when an individual used them to describe an interaction

Knowledge, Practice, and Motivation Study participants' understandings of knowledge and practice corresponded with the definitions in the user manual, but their understandings of motivation were not nearly as well aligned with the manual definitions. Specifically, when describing how an interaction that they planned to enter in their logs was related to these concepts, participants consistently matched the manual definitions for knowledge (88%) and practice (88%) but not nearly as often for motivation (63%).

When asked in cognitive interviews, study participants indicated understandings of knowledge that matched the definition in the log manual 95% of the time. The following three responses—from a math specialist, a literacy specialist and a principal respectively—are representative:

- Knowledge is basically if they made me think about something in a different way or if I learned something different. (October 19, 2005)
- Knowledge I tend to think of as their specific content area maybe background knowledge. Knowledge for the standards, knowledge of theory, philosophy. (October 20, 2005)
- Knowledge is what you know about a particular subject or a particular area... It gets kinda case specific as far science, social studies, reading, language arts. And... when it's in reference to subject matter it's your knowledge of the subject

matter. When it's about a particular student it's from being in a school, it's your knowledge of that particular student. It's just what you know about a particular thing or person. (October 19, 2005)

These participants' understandings of knowledge not only corresponded with the log manual but also covered various types of knowledge, including that of subject matter, students, and standards or curricula.

Participants' understanding of practice matched the log manual 85% of the time. The following responses, from a literacy specialist and a mathematics specialist, are representative:

- Practice is about pedagogy; you know the methods that they're using. (October 20, 2005)
- Practice is doing; you know, actually doing things. Did it make me change the way I do things. .. or am I trying to change the way they do things? (October 19, 2005)

With respect to motivation, however, study participants' understanding corresponded with the log manual much less. When asked to define motivation in cognitive interviews, 90% gave definitions that corresponded with the manual. However, when participants reported an interaction as one that influenced motivation, their understanding of motivation matched the LDP log user manual for only 63% of the interactions. Where participants' understanding matched the user manual, the interactions focused on their motivation or that of another staff member.

When their understanding of motivation did not correspond to the manual, study participants often linked it to student motivation rather than to their own motivation or to a colleague's. This poses a problem in that the log attempts to get participants to distinguish between an interaction intended to influence their motivation, knowledge, or practice or that of a colleague.¹⁰ For example, a reading specialist described an interaction that she had with a reading teacher after observing her teach a vocabulary lesson:

I would like to think it was about all three. Giving [the reading teacher] some knowledge in good vocabulary instruction which hopefully would impact her practice and she'd stop doing that [having students look words up in the dictionary]. And then hopefully then that would *motivate students* to like to learn the words better. To motivate them more than, dictionary is such a kill and drill. (October 20, 2005; italics added for emphasis)

Although the participant's description of this interaction suggests that her understanding of knowledge and practice is consistent with that of the LDP log user manual, her understanding of motivation is not; that is, it focused on student motivation rather than on teacher motivation. We are not questioning the accuracy of the

¹⁰As noted earlier, in this pilot study of the LDP log, we did not include interactions with students and parents, although we acknowledge that students are important to understanding leadership in schools (see Ruddock, Chaplain, & Wallace, 1996). Our redesigned log includes interactions with parents and students.

reading specialists' account; rather, what is striking us is how she understands motivation entirely in terms of student motivation.

For about half the nonmatching cases (i.e., nine interactions across six participants), study participants referred to motivation in terms of motivating students rather than themselves or colleagues. In describing three more interactions, study participants referred to both student and teacher motivation. For example, a mathematics teacher enlisted a science teacher to help teach a mathematics lesson and described how this interaction influenced knowledge, practice, and motivation:

And *motivation*, when you show a child you know when you can get a child to become in touch with their creative side they just, they become really motivated and the teachers become motivated by watching how motivated the students are. (October 20, 2005)

This example points to a larger issue; it highlights how influence is often not direct but indirect: An influence on a teacher's knowledge and practice can in turn result in changing students' motivation to learn, which can in turn influence a teacher's motivation to teach. Logs of practice may be crude instruments when it comes to picking up the nuances of influence on motivation.

Direction of Influence The LDP log required participants to select a direction of influence for each interaction that they logged; that is, either a participant attempted to influence someone else (i.e., provide information), or someone or something else attempted to influence the participant (i.e., solicit information). In cases where several topics were discussed in one interaction, participants are asked to "please consider who initiated the interaction." Our analysis suggests that this item was especially problematic, given that low levels of correspondence between participants' understanding and the manual.

Two thirds of the participants reported that they struggled to select a direction of influence. For approximately 25% of the interactions ($n = 26$) described in the cognitive interviews, participants reported that the direction of influence went both ways in that they intended to influence a colleague (or colleagues) and that they themselves were influenced. For example, a principal described an interaction that involved checking in with teachers in their classrooms, where the influence was bidirectional. In this interaction (as described by the principal), a teacher shared her plans for reading instruction, and the principal made suggestions about how the teacher could make it both a reading and a writing activity. When asked about the direction of influence, the principal reported, "I think initially the attempt was to influence me. But, as I provided the activities for her to have, I think I ended up being the influential party" (October 28, 2005). Participants identified no direction of influence in only 4 of the 97 interactions.

Planned or Spontaneous In discussing their log entries, over half the study participants (13 participants across 22 interactions), struggled with choosing whether an interaction was planned or spontaneous. Interactions that some participants consider planned, others considered spontaneous. Furthermore, participants expressed difficulty in their designation because part of an interaction might be planned whereas another part might be spontaneous.

Participants identified 12 of 99 total interactions as being both planned and spontaneous, thus making it difficult for them to choose an option in the LDP log. These interactions tended to start with something planned, but then the aspect of the interaction that they discussed became spontaneous. For example, a literacy specialist described helping a mathematics teacher:

This one I have to think about. It was a planned to visit him, but it was spontaneous to see the flaw and try and fix it. So I would say that I'm going to mark spontaneous but it was within a planned [visit], I was supposed to come this morning to see him. (October 20, 2005)

The literacy specialist's statement captures the difficulty of distinguishing a planned meeting from the spontaneity of the substance that emerged within the interaction.

In nine of the interactions described in cognitive interviews, participants reported struggling with deciding whether a generally planned interaction was planned or spontaneous. Participants were aware that the interaction would occur, even though there was no allotted time for the interaction. In some instances, the general time of the interaction was known in advance, but neither the topic nor the location was planned. For example, a mathematics teacher described an informal meeting that occurred with a colleague every morning:

It's difficult to say because we meet everyday even though we're supposed to meet twice a week we literally meet everyday; we don't start our day without talking to each other about something before the students come in. So I would kinda say at this point it's planned because it would be weird if we didn't talk before the students came in. (October 26, 2005)

For this participant, this interaction occurred regularly; thus, it was planned. However, according to the participant's interpretation of the user manual definition, the interaction was technically spontaneous because the subject, time, and location of the interaction were not predetermined.

Participants described nine interactions as being difficult to define as planned or spontaneous, namely, because the interaction was planned for one person and spontaneous for the other. An assistant principal, for example, described an interaction in which she followed up with the two lead literacy teachers in the school about their experience working with teachers to implement a new strategy in their classrooms:

It was planned. The specific time wasn't planned but I knew today was gonna be the first day so I wanted to make sure that I had an opportunity to touch base with the teachers to see how this particular interaction went with the teachers because they have been challenged with some of the staff members. (October 24, 2005)

From the perspective of the two literacy teachers, the interaction was not planned; from the assistant principal's perspective, however, it was planned. Whether something is planned or spontaneous does indeed depend on whom one asks in an interaction.

Our analysis of the cognitive interview data underscores the fuzzy boundary between planned and spontaneous interactions. In particular, these accounts underscore the emergent nature of interactions. Although an interaction might start out as planned from the perspective of at least one participant, it becomes spontaneous

because of the emergent nature of practice. Furthermore, what it means for something to be planned for school staff does not necessarily mean *scheduled* in terms of time and place but merely that staff members plan to do something, sometime during that day. For example, two administrators described keeping running lists in their heads of things to do that they would get to when there was a free moment or when it became necessary. These interactions could easily fall into the spontaneous or planned category in the LDP log.

9.5.3 Research Question 3

To what extent do study participants and the researchers who shadowed them agree when using the LDP log to describe the same social interaction?

Concurrent Validity: Comparing Log Data and Observer Data Although our analysis to this point surfaces some important issues with respect to study participants' understandings of key terms, we found high agreement between LDP log data and the shadowing data generated by observers. Agreement between the LDP log and the shadowing data was high, 80% or above for all categories (see Appendix E), thereby suggesting that the log accurately captures key dimensions of leadership practice as experienced by study participants on the data collection days. Agreement was highest (94.4%) for the time of the interaction (see Table 9.6), which is noteworthy because study participants did not complete their logs until the end of the day. With respect to who the interaction was with or what it was about, study participants and observers agreed for 88.4% of the interactions. For how the interaction occurred, the logger and observer responses were a 86.3% match.¹¹ Regarding where the interaction took place, 80.6% of the interactions were a match. With respect to what happened in an interaction, agreement was 85.1%.¹²

All kappa coefficients were statistically significant at the .001 level (see Table 9.7). The highest agreement between log and shadow data involved the time

Table 9.6 Logger and observer reports: percentage match of interactions

	What	Who	Where	How	Time ^a
Match	85.1	88.4	80.6	86.3	94.4
No match	14.9	11.6	19.4	13.7	5.6

Note: Number of interactions varied across categories, from a high of 71 (time) to a low of 51 (how)

^aBefore school, 9 a.m. to noon, noon to 3 p.m., and after school

¹¹The logging instrument collected how the interaction occurred in cases where the interaction occurred with an individual and not with a group or resource (51 out of 71 total individual interactions).

¹²This calculation used the conservative decision rule, whereby if a participant's log entry was too vague to verify, then this response was counted as a nonmatch.

Table 9.7 Kappas of logger–shadower interactions

	Where	How	Time ^a
<i>n</i>	67	51	71
Kappa	.758	.711	.915
<i>SE</i>	.0568	.0894	.0814
Agreement (%)	80.60	86.37	94.37

Note: All kappa coefficients are significant at the $p < .001$ level

^aTime: before school, 9 a.m. to noon, noon to 3 p.m., and after school

of day that the interaction occurred, with a kappa coefficient of .915. The location of the interaction was on the border between being an excellent and a good measure of validity, with a kappa of .758. Although agreement was not as strong, how the interaction occurred was still a good measure of reliability, with a kappa coefficient of .7111.

9.5.4 Research Question 4

How representative are study participants' log entries regarding the types of social influence interactions recorded by researchers for the same logging days?

Selection Validity: Are Study Participants' Log Selections Biased? Contrary to our expectations, our findings revealed few significant differences in the characteristics of logged interactions as compared to the larger sample of interactions recorded by observers on the same days—our approximation for the population of interactions (see Table 9.8). There were no significant differences between study participants and observers in the number of interactions reported at specific times of the day (e.g., early morning, late afternoon). Furthermore, there were no significant differences between the focus of the interaction as reported by study participants and observers. Across the remaining characteristics—where, how, and with whom an interaction took place—there were some significant differences between the types of interactions that study participants reported and the interactions as documented by observers.

There were a handful of categories in which the interactions captured by the LDP log differed from our approximation for the population of interactions as captured by the observers, thereby raising the possibility that study participants may be more likely to select interactions with particular characteristics for inclusion in the LDP log (see Table 9.8).¹³ First, our analysis suggests that study participants may be disposed to select interactions outside their own offices and less likely to pick interactions that happen within them. Second, study participants undersampled

¹³Note that study participants were much less likely to report mathematics interactions, as opposed to interactions dealing with other subjects. However, this is not a statistically significant difference.

Table 9.8 Comparing shadower and logger populations of interactions in all schools

	Shadower		Logger		Difference L% – S%
	<i>n</i>	%	<i>n</i>	%	
Time					
Before 9 a.m.	41	25.8	25	28.1	2.3
9:00–11:59 a.m.	62	39.0	32	36.0	–3.0
12:00–2:59 p.m.	52	32.7	28	31.5	–1.2
3 p.m. or after	4	2.5	4	4.5	2.0
Total	159	100.0	89	100.0	
Where					
My office	43	27.2	12	14.3	–12.9**
Main office	15	9.5	8	9.5	0.0
Classroom	60	38.0	28	33.3	–4.6
Staff room	2	1.3	4	4.8	3.5
Conference room	1	0.6	5	6.0	5.3*
Hallway	21	13.3	16	19.1	5.8
Other location (e.g., library, cafe)	16	10.1	11	13.1	3.0
Total	158	100.0	84	100.0	
How					
Face-to-face: One-on-one	102	65.0	50	74.6	9.7
Phone/intercom	6	3.8	1	1.5	–2.3
E-mail/internet	6	3.8	1	1.5	–2.3
Document/book	21	13.4	3	4.5	–8.9*
Face-to-face: Small group (2–5)	18	11.5	8	11.9	0.5
Face-to-face: Large group (6+)	4	2.6	4	6.0	3.4
Total	157	100.0	67	100.0	
Subject^a					
Mathematics	73	47.4	28	37.3	–10.1
Reading	47	30.5	21	28.0	–2.5
English/language arts (+ writing)	5	3.3	4	5.3	2.1
Science	2	1.3	2	2.7	1.4
Multiple subjects	13	8.4	9	12.0	3.6
Other subject (arts, music, other)	13	8.4	10	13.3	4.9
Social studies	1	0.7	1	1.3	0.7
Total	154	100.0	75	100.0	
With whom^b					
Principal	11	7.1	11	13.9	6.9
Assistant principal	6	3.9	2	2.5	–1.3
Math specialist	10	6.4	3	3.8	–2.6
Literacy specialist	4	2.6	4	5.1	2.5
Teacher (includes special ed)	85	54.5	46	58.2	3.7
Other in school (e.g., other staff)	9	5.8	4	5.1	–0.7
Other outside of school	8	5.1	5	6.3	1.2
Materials (curriculum, text, documents)	23	14.7	4	5.1	–9.7**
Total	156	100.0	79	100.0	

* $p < .05$. ** $p < .01$ ^aCoded as math if multiple subjects included math^bIf multiple people, then counted only the person with highest status, defined by list order

interactions that involved inanimate objects (e.g., book, curricula) and overreported formal interactions (e.g., meetings) and face-to-face interactions. Overall, comparing the characteristics of the interactions logged by study participants to the characteristics of all interactions recorded by observers—our approximation for the population of interactions—suggests that with a few exceptions, loggers are relatively unbiased in selecting from the range of interactions in which they engage as related to mathematics and/or curriculum and instruction.^{14,15}

9.6 Discussion: Redesigning the LDP Log

The purpose of our study was to examine the validity of the inferences that we can make based on the LDP log data with respect to what actually happened to study participants, to redesign the LDP log. We consider the entailments of four issues that our analyses surfaced in terms of redesigning the LDP log.

One issue involves sampling—that of logging days and that of interactions within days. To use the LDP log to generalize leadership practice across a school year, we need a sampling strategy that taps into the variation in leadership across the school year. One response might be to sample days from a school year at random. However, a random sampling strategy does not take into account critical events and seasonal variation in leadership practice (e.g., start of year events), and it may not pick up on events that happen monthly or quarterly or that structure leadership interactions in schools. A stratified sampling strategy targeting a couple of weeks at different times of the school year seems necessary to pick up on seasonal variation. With respect to sampling interactions within days, a key issue to consider in redesigning the LDP log is whether to allow participants to select social interactions from across the day, instead of one interaction per hour. Our analysis suggests that for some school leaders—especially, leaders (formally designated or informal) who have full- or part-time classroom teaching responsibilities—social influence interactions are unevenly distributed across the school day. Hence, a sampling strategy that requires study participants to sample one interaction per hour may miss key social influence interactions that are concentrated in particular times in the day when such leaders are not teaching.

A second issue concerns a different sort of sampling—namely, study participants' selection of interactions to log. Specifically, we need to consider how to minimize study participants' sampling bias through training and through the redesign of the LDP log user manual. For example, stressing that interactions with inanimate objects (e.g., curriculum materials) are important in social influence

¹⁴Note that the small sample size in some cases affects the detection of significant differences. In cases where a relatively large difference exists but is not significant, we make an effort to highlight it.

¹⁵A detailed description of the validity and reliability of the Experience Sampling Method log is beyond the scope of this article. For more information see Konstantopoulos, 2008.

interactions might help reduce the tendency for study participants to undersample these types of interactions.

A third issue that our analysis surfaced with respect to redesigning the LDP log—including the user manual and prestudy training sessions—concerns some of the terms used to characterize social influence interactions and the options available to participants. First, a clearer and more elaborate description of motivation is necessary, with specific reference to teacher and administrator motivation. Our analysis suggests that motivation is often indirect and that discussion of direct and indirect motivation might help participants become aware of different ways in which motivation might work—for example, changes in teaching practice motivate students, which in turn motivates a teacher. Second, our analysis suggests that in redesigning the LDP log, we will need to expand the options under direction of influence to allow for bidirectional influence. Furthermore, the wording of the direction-of-influence question—with its focus on (a) providing information or advice and (b) soliciting and receiving information or advice from a colleague—appears to confuse rather than clarify the direction-of-influence issue. Moreover, we will need to separate direction of influence from who initiates the interaction.

A third and more difficult redesign challenge concerns getting participants to distinguish the intent to influence from actually being influenced. From our perspective, the intent to influence someone or be influenced is sufficient for defining that interaction as a leadership activity. Whether the interaction actually influenced an individual's motivation, knowledge, and/or practice is a related but different matter—it concerns the efficacy of the leadership activity. A fourth design challenge involves reworking the question that attempts to distinguish spontaneous from planned interactions. The user manual and training can be redesigned such that participants are directed to decide whether something is planned or spontaneous from their perspective rather than from the perspective of other participants in the interaction. A somewhat more difficult redesign decision concerns which dimensions of an interaction should be used to determine whether an interaction is planned or spontaneous, such as the timing or the place.

A fourth issue that our analysis surfaced concerns whether and how the LDP log might be redesigned so that it can situate particular interactions in a broader context. One possibility is to include an open-ended item that asks loggers to reflect on how each interaction they log connects with their personal and professional goals, thereby embedding the interaction in a broader context. Letting study participants enter into the log information that they think relevant to the interaction could generate data that would allow the interaction to be situated in a broader context. In this way, the LDP log could capture the logger's perspective. The decision to include such an open-ended item, however, must take into account the extra response burden that such items place on study participants. As a math specialist put it, the closed-ended items make it easy on respondents “because a lot of it is fill-in. .. and that of course makes it very easy” (October 28, 2005). The LDP log—indeed, logs in general—may not be the optimal methodology for getting at the underlying professional and personal meanings and goals of those participating in social influence interactions. Although logs are good at capturing the here and now, they are not optimal for capturing how events in the past structure and give meaning to current

practice. Hence, an alternative strategy might combine the LDP log and in-depth interviews with a purposeful subsample of study participants to collect data that would help situate interactions within participants' personal and professional goals. Moreover, analysis of log data could be the basis for purposefully sampling participants and for grounding interviews with them.

9.7 Conclusion

The LDP log provides a methodological tool for studying school leadership practice in natural settings through the self-reports of formally designated leaders and informal school leaders. This article reports on the validity of the data generated by the LDP log. Analyzing a combination of log data, observer data, and data from cognitive interviews—based on a triangulation approach—we examined the validity of leadership practice as captured by the LDP log. Overall, we found high levels of agreement between what study participants reported and what observers recorded (based on their observations of study participants). Furthermore, in comparing all the interactions documented by observers for days in which school leaders made log entries, we found that (with few exceptions) the patterns captured in the log were similar to those found in the shadow data. In other words, study participants' sampling decisions were, for the most part, not biased in favor of some types of interactions over others. Although the LDP generates robust data (with some important exceptions discussed above), our analysis suggests that a key concern involves sampling of days and interactions within days. Moreover, we need to work on rethinking how we present some key descriptors of interactions in the log, manual, and study participants' training.

As a research methodology, logs in general and the LDP log in particular enable us to gather data on school leadership practice across larger samples of schools and leaders (formally designated and otherwise) than what is possible with the more labor-intensive ethnographic and structured observation methodologies. Although the LDP log is more costly to administer than school leader questionnaires, it generates more accurate measures of practice because of its proximity to the behavior being reported on. Research shows that annual surveys often yield flawed estimates of behaviors because respondents have difficulty accurately remembering whether and how frequently they were engaged in a behavior (Tourangeau et al., 2000). Because the LDP log is completed daily, it reduces this recall problem. Although the LDP log has limitations, it can be a valuable tool for gathering information on large samples of schools and leaders, which is critical in efficacy studies of leadership development programs. Moreover, our intent is not to suggest that the LDP log or any other log methodology should supplant existing surveys or ethnographic studies of leadership practice that dominate the field. Rather, our intention is to develop and study an alternative methodology that can supplement existing methods, which is critical if we want to generate robust empirical data critical for large sample and efficacy studies.

Appendices

Appendix A: Daily Practice Log

Instructional Leadership Daily Practice Log

Name:

School:

Was this a typical day?

Yes

No If No, Please explain:

Approximately how many interactions did you have today that were related to mathematics or curriculum & instruction?

Did you complete the interaction chart throughout the day?

Yes

No

Please estimate what percentage of your day was spent on the following tasks? (i.e. 15%)

<p>School / Department/ Grade Administrative Duties <i>i.e. Managing budgets, resources, personnel, schedules, facilities, students and school improvement.</i></p>	<input style="width: 50px;" type="text"/> %
<p>Curriculum & Instructional Leadership Duties <i>i.e. Observing students or teachers; reviewing lesson plans, student work, curricula, test data; discussing teaching practice, curricula, professional development, assessment.</i></p>	<input style="width: 50px;" type="text"/> %
<p>Classroom Teaching Duties <i>i.e. Duties associated with classroom teaching. i.e. preparing lesson plans or classroom materials, grading student work.</i></p>	<input style="width: 50px;" type="text"/> %
<p>Non-Teaching Duties <i>i.e. Activities not directly tied to teaching i.e. collecting money or permission forms, supervising playground or lunchroom.</i></p>	<input style="width: 50px;" type="text"/> %
<p>Other</p> <input style="width: 100%; height: 20px;" type="text"/>	<input style="width: 50px;" type="text"/> %

Sum of percentages should not exceed 100.

During the following time periods, did you try to influence a colleague(s)' knowledge, practice, or motivation related to mathematics or curriculum & instruction



OR

Did a colleague(s) or resource influence your knowledge, practice, or motivation related to mathematics or curriculum & instruction?

Select yes or no for each time period.
When you select yes, you will be prompted to describe this interaction. When you are finished with all time periods, please click on the button below to submit your answers.

7:00 - 7:59 am	<input type="checkbox"/> No	<input type="checkbox"/> Yes	12:00 - 12:59 pm	<input type="checkbox"/> No	<input type="checkbox"/> Yes
8:00 - 8:59 am	<input type="checkbox"/> No	<input type="checkbox"/> Yes	1:00 - 1:59 pm	<input type="checkbox"/> No	<input type="checkbox"/> Yes
9:00 - 9:59 am	<input type="checkbox"/> No	<input type="checkbox"/> Yes	2:00 - 2:59 pm	<input type="checkbox"/> No	<input type="checkbox"/> Yes
10:00 - 10:59 am	<input type="checkbox"/> No	<input type="checkbox"/> Yes	3:00 - 3:59 pm	<input type="checkbox"/> No	<input type="checkbox"/> Yes
11:00 - 11:59 am	<input type="checkbox"/> No	<input type="checkbox"/> Yes	4:00 - 4:59 pm	<input type="checkbox"/> No	<input type="checkbox"/> Yes

I have completed all the time periods.



NORTHWESTERN UNIVERSITY

*Thank you for completing this Daily Practice Log.
Please click "Send" when you are done.*

SEND

Please be patient while waiting for the data to be sent.
A confirmation message will indicate when the data has been successfully sent.

Appendix B: Document That Observer’s Used to Record/Input Data while Shadowing

DAY		DATE		TEACHER / ADMINISTRATOR			SCHOOL		INIT	
I / B	TIME	WHERE	WITH WHOM / WHAT #	HOW	SUBJ	WHAT’S HAPPENING	CODE	PLANNED / SPONTANEOUS	PROVIDE SOLICIT	

Code: 01 – ADMIN 02 – c + I LEADERSHIP 03 – CLASSROOM TEACHING 04 – NON-TEACHING 05 – OTHER

Appendix C: Sample of the Cognitive Interview – Post-Logging Protocol

The goal of this interview is for researchers to understand your thinking when completing the daily practice log. We would like you to share with us how you will enter these interactions into the daily practice log and to explain your decision making process.

(1)

- (a) The log asks you to determine if an interaction influenced your knowledge practice or motivation, how would you define EACH of these terms?

Knowledge, Practice, Motivation

For the next set of questions please reflect on the THREE interactions that are most closely tied to mathematics or curriculum & instruction that intend to enter in the daily practice log.

You will need to REPEAT questions 2–7 for each of their three interactions most closely tied to mathematics or curriculum & instruction.

(2)

- (a) Regarding your [Insert the name or description of the interaction] interaction, when did it take place and who or what did it occur with?
- (b) How will you rank this interaction on the influence scale?

Not influential, Somewhat influential, Influential, Very influential, Extremely influential Why did you give this interaction that ranking?

- (3) Would you consider this interaction to be an example of mathematics leadership? (*The participant may ask what we mean by math or curriculum & instruction leadership, but we are interested in what they consider leadership to be.*)

How is this leadership for mathematics?

OR

(*If the interaction was Not related to math*)

- (a) Would you consider this interaction to be an example of curriculum & instruction leadership? (*The participant may ask what we mean by math or curriculum & instruction leadership, but we are interested in what they consider leadership to be.*)

How is this leadership for curriculum & instruction?

- (4) Did you influence a colleague(s) or did a colleague(s) or resource influence you. Depending on response? How did you decide [mention response]?
- (5) How did you decide this interaction was [insert response – spontaneous or planned]?
- (6) How was this interaction about [include response – knowledge, practice or motivation].
- (7) *Ask this question only if the interaction pertained to math* Regarding this interaction, please explain from which of the following did this math interaction stem from? Student textbook, teacher's guide, other curricular materials, student comment or response, student written work, assessment materials, standardized tests, standards documents or other.
- (8) Did you use the interaction chart throughout the day? Was this tool useful when you entered the information into the daily practice log at the end of the day? Please explain.
- (9) On a scale of 1–10 (1 being easy, 10 being extremely difficult) how difficult was it to use the interaction chart? Please explain.
- (10) On a scale of 1–10 (1 easy, 10 being extremely difficult) how difficult has it been to complete the daily practice log?
- (11) Approximately how long has it taken to complete the log each day?
- (12) After completing the daily practice log do you find that it accurately captures the nature of your interactions about mathematics or curriculum & instruction for the day? Please explain.
- (13) After completing the daily practice log do you find that it accurately captures the leadership for mathematics or curriculum & instruction as you experience it for
- (a) The day? If so, how? If not, how not?
- (b) In this school this year? If so, how, if not, how not?
- (14) Is there anything else that you would like to share about your experience completing the daily practice log?

Additional/Recorded Questions from the 2nd Round of INT

- (10) Do you have any recommendations on what could be done to improve the process of completing the daily practice log? (e.g. Would they prefer a paper copy or to email their results)
- (11) On a scale of 1–10 (1 very uncomfortable, 10 being extremely comfortable) how would you describe your level of comfort with computers & technology?
- (12) On a scale of 1–10 (1 unskilled, 10 being extremely skilled) how would you describe your skill level with computers?
- (13) *Participants may not be able to answer all of these questions.*
 - (a) From which location did you most frequently complete the log? (e.g. home, classroom, library, office)
 - (b) What type of computer is this? (e.g. PC or Mac)
 - (c) What is the processing speed? (e.g. Pentium II/III or Powerbook G3/G4)
 - (d) What operating system does this computer have? (e.g. Windows XP, NT, 2000, 1998 or OS 8, 9, 10)
 - (e) What type of internet connection does this computer have? (e.g. dial-up, DSL, T1, cable modem)
 - (f) What type of browser does this computer have? (e.g. Internet Explorer, Netscape, Mozilla, Foxfire)
- (14) After completing the daily practice log do you find that it accurately captures the nature of your interactions about mathematics or curriculum & instruction for the day? Please explain.
- (15) After completing the daily practice log do you find that it accurately captures the leadership for mathematics or curriculum & instruction as you experience it for
 - (a) The day? If so, how? If not, how not?
 - (b) In this school this year? If so, how, if not, how not?
- (16) Do you have any recommendations on how researchers can capture and best study instructional leadership at the school level?
- (17) Is there anything else that you would like to share about your experience completing the daily practice log?

Appendix D: Inter-rater Reliability Across Observers

As a check on reliability, two members of the fieldwork team observed one participant during one day of the study. The data was entered into a database under the same topical structure as the data collection form. Then the data from both observers was matched by interaction, resulting in pairs of observations. The observations where there was no corresponding data for the interaction from the other observer were left single.

The observations were matched by first looking at the time to see if they were similar and then examining the location and who was participating in the interaction. If both were similar then this was considered a match. Thus, if the time or what took place were not similar this was left as a single un-matched interaction. The most conservative approach was taken towards matching these pairs of observations such that if the observations did not provide an exact match, this was not evaluated as a match. A total of 32 interactions were compared.

The N for the % matches is based on the total number of interactions recorded by both observers during the day. This means that if one observer recorded an interaction, but the other observer did not, then this is included in the N. This occurred three times for each observer, resulting in a total of 6 interactions. A non-match (or 0) is scored for each of these interactions since no record indicates a lack of agreement. Thus, the highest level of agreement possible in any category is 32 out of the total 38 interactions (or 82.4%).

Next, kappa coefficients were calculated to provide an additional and stronger test for reliability. To calculate a kappa, we coded the data into discrete categories. The categories for Where, How, and the Time of the interaction were assigned numerical codes (see Appendix F for exact codes). Observers recorded exactly what time the interaction occurred (hour and minute), so codes were assigned to designate whether the interaction occurred roughly before school (before 9 am), in the morning of the school day (9 am–11:59 am), during school in the afternoon (12 pm–2:59 pm) and roughly after school (3 pm and after). Kappa coefficients were calculated for these four categories (What Activity Type, Where, How, and Time) using the kappa function in the statistical program STATA (see Appendix G for an example of how to calculate a kappa coefficient). Two categories – What Happened and With Whom the interaction took place – proved difficult to calculate kappa statistics due to the descriptive nature of the categories. Specifically, “who” the interaction took place with became too complex to code both because of the multitude of people the interactions took place with, but also because the interactions often took place with more than one person, making it difficult to even categorize by role within school. Thus, no kappa coefficients are calculated here.

Results. Overall, the agreement between the two observers was high with respect to what the shadowed study participant was doing and the high kappas indicate agreement that cannot be attributed to chance. We found that the two observers agreed on where the interaction took place for 81.6% of the interactions how the interaction occurred for 79.0% of the interactions (see Table 9.9). The exact time recorded by each observer also matched for 81.6% of the interactions. Just slightly less, 79.0% of agreement was found for how the interaction occurred. Observers

Table 9.9 Double-shadower percent matches of interactions

	What	Who	Where	How	Time
Match	76.3%	71.1%	81.6%	79.0%	81.6%
No match	23.7%	28.9%	18.4%	21.1%	18.4%

N = 38 interactions; includes all interactions that at least one shadower recorded

Table 9.10 Kappas of double-shadower reports of interactions

	Where	How	Time
(N)	32	32	32
Kappa	0.929	0.889	1.000
(Std Error)	(.1763)	(.1392)	(.1280)
Prob>Z	0.0000	0.0000	0.0000
Agreement (%)	96.88%	93.75%	100.00%

matched descriptions of what was happening in 76.3% of the interactions and agreed 71.1% of the time about with whom or what the interaction occurred. It should be noted that this percent match might be low as a result of observer error in recording who the interaction occurred with – especially early on in shadowing when the observer did not know everyone.

Kappa coefficients were calculated using the 32 interactions that both observers recorded. For these 32 interactions, the resulting kappa coefficients were all statistically significant suggesting high reliability (see Table 9.10). The time of the interaction, as coded into part of the day, had a kappa coefficient of 1. Where the interaction took place had a kappa coefficient of .929, and how it occurred had a kappa coefficient of .889. These high kappas show that the information collected over categories by different observers recording the same interaction is quite consistent. However, the coefficients do not account for the three interactions that each observer recorded which the other did not. Still, this only affected 3 (or 8.5%) of the total thirty-five interactions recorded by each observer.

Appendix E: Examples of Matches in Logger/Shadow Interactions

What:

Match (=1)

Logger: It was a planned interaction for me to be in Larry's room and working with Literature circles with his students. I noticed during this time that it wasn't working as well as I would have liked with this class.

Shadower: Ms. R is observing Mr. P's classroom and helping w/his 'literacy circles.' Ms. R goes over the 'expectations' of working in small groups.

Vague Match (=1) [note: there were 7 vagues out of 64 matches].

Logger: I need to find out more details about upcoming math inservices.

Shadower: Mrs. F left a message for Dr. Long regarding math professional development sessions. [Next interaction – with computer – is: Mrs. F tries to find Dr. Long's CPS email address in order to contact him. A teacher assists her in finding this address.]

No Match (=0)

Logger: social worker wanted students to be notified that if they write anything about harming themselves in their journal, he will have to report it.

Shadower: (At staff meeting) they discuss suicidal student and the new person in charge of the boys program.

Who:

Match (=1)

L: Principal

S: Principal

Vague Match (=1)

L: Internal Walk-through team

S: art teacher, library specialist, and principal

OR

L: Mr. Humbert (teacher)

S: teacher

No Match (=0)

L: my internal walk-through team; co-leader: Ms. Damlich Ms. Freeman Ms. Ryder

S: two teachers

Time:

Match (=1) anytime within the shadower’s hour (12:00–12:59)

L: 12:34

S: 12:45

No Match (=0)

L: 12:34

S: 1:10

Appendix F: Codes Used to Calculate Kappa Coefficients

Codes for Kappas:

How:

1 = Face to Face: one on one

2 = phone / intercom

3 = email / internet

4 = document / book

5 = Face to Face: small group (2–5)

6 = Face to Face: large group (6+)

Where:

1 = My office

2 = Main office

3 = Classroom

- 4 = Staff room
- 5 = Conference room
- 6 = Hallway
- 7 = Other location in school (library, cafeteria...)

Time:

- 1 = before 9am (Before school day)
- 2 = b/w 9–11:59 am (AM school day)
- 3 = b/w 12–2:59 pm (PM school day)
- 4 = 3pm or after (After school day)

School (pseudonyms):

- 1 = Acorn
- 2 = Alder
- 3 = Ash
- 4 = Aspen

Logger Role:

- 1 = Principal
- 2 = Asst. Principal
- 3 = Specialist
- 4 = teacher

Appendix G: Example of Calculating the Kappa Coefficient

1. Matrix Comparing Observer 1 to Observer 2 Recordings

How 1	How 2						Total
	A	B	C	D	E	F	
A. Face to face: One on one	33	1	0	0	4	0	38
B. Phone/intercom	0	1	0	0	0	0	1
C. Email/internet	0	0	1	0	0	0	1
D. Document/book	0	0	0	3	0	0	3
E. Face to face: Small group (2–5)	1	0	0	0	5	0	6
F. Face to face: Large group (6+)	0	0	0	0	1	1	2
Total	34	2	1	3	10	1	51

2. Calculate q; the # of cases expected to match by chance

$q = n(\text{row}) * n(\text{col})/N$	
A =	25.33333
B =	0.039216
C =	0.019608

D =	0.176471
E =	1.176471
F =	0.039216
q = total =	26.78431

3. Calculate Kappa

Kappa = $(d - q)/(N - q)$	
d = diagonal total =	44
N = total =	51
[if match = 100%, d = N]	
Kappa =	0.71

References

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, UK: Cambridge University Press.

Barnard, C. (1938). *The functions of the executive*. Cambridge, MA: Harvard University Press.

Bass, B. (1990). *Bass and Stogdill's handbook of leadership: Theory, research, and managerial applications*. New York, NY: Free Press.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*, 579–616.

Bossert, S. T., Dwyer, D. C., Rowan, B., & Lee, G. V. (1982). The instructional management role of the principal. *Educational Administration Quarterly, 18*, 34–64.

Camburn, E., & Barnes, C. (2004). Assessing the validity of a language arts instruction log through triangulation. *Elementary School Journal, 105*, 49–74.

Camburn, E., & Han, S. W. (2005, April). *Validating measures of instruction based on annual surveys*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Camburn, E., Rowan, B., & Taylor, J. E. (2003). Distributed leadership in schools: The case of elementary schools adopting comprehensive school reform models. *Educational Evaluation and Policy Analysis, 25*(4), 347–373.

Camburn, E., Spillane, J. P., & Sebastian, J. (2006, April). *Measuring principal practice: Results from two promising measurement strategies*. Paper prepared for presentation at the annual meeting of the American Educational Research Association, San Francisco, CA.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of experience sampling method. *The Journal of Nervous and Mental Disease, 175*, 526–536.

Cyert, R. M., & March, J. G. (1963). *A behavioral theory of the firm*. Englewood Cliffs, NJ: Prentice Hall.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods* (2nd ed.). New York, NY: McGraw-Hill.

- Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Eccles, R. G., & Nohria, N. (1992). *Beyond the hype: Rediscovering the essence of management*. Boston, MA: Harvard Business School Press.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York, NY: Wiley.
- Gorin, A. A., & Stone, A. A. (2001). Recall biases and cognitive errors in retrospective self-reports: A call for momentary assessments. In A. Baum, T. Revenson, & J. Singer (Eds.), *Handbook of health psychology* (pp. 405–413). Mahwah, NJ: Erlbaum.
- Gronn, P. (2000). Distributed properties: A new architecture for leadership. *Educational Management and Administration*, 28(3), 317–338.
- Gronn, P. (2002). Distributed leadership. In K. Leithwood & P. Hallinger (Eds.), *Second international handbook of educational leadership and administration* (pp. 653–696). Dordrecht, The Netherlands: Kluwer.
- Gronn, P. (2003). *The new work of educational leaders: Changing leadership practice in an era of school reform*. London, UK: Chapman.
- Hallinger, P., & Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980–1995. *Educational Administration Quarterly*, 32(1), 5–44.
- Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *Elementary School Journal*, 86(2), 217–247.
- Heck, R. H., & Hallinger, P. (1999). Next generation methods for the study of leadership and school improvement. In J. Murphy & K. Louis (Eds.), *Handbook of educational administration* (pp. 141–162). New York, NY: Longman.
- Heifetz, R. A. (1994). *Leadership without easy answers*. Cambridge, MA: Belknap Press.
- Heller, M. F., & Firestone, W. A. (1995). Who's in charge here? Sources of leadership for change in eight schools. *Elementary School Journal*, 96, 65–86.
- Hilton, M. (1989). A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. *British Journal of Addiction*, 84, 1085–1092.
- Hollander, E. P., & Julian, J. W. (1969). Contemporary trends in the analysis of leadership processes. *Psychological Bulletin*, 71, 387–397.
- Katz, D., & Kahn, R. L. (1966). *The social psychology of organizations*. New York, NY: Wiley.
- Konstantopoulos, S. (2008, April). *Validity and reliability of Experience Sampling Methods (ESM) in measuring school principals' work practice*. Paper presented at the annual American Educational Research Association, New York, NY.
- Lakomski, G. (2005). *Managing without leadership: Towards a theory of organizational functioning*. London: Elsevier.
- Leithwood, K., Seashore-Louis, K., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning: A review of research for the learning from leadership project*. New York, NY: Wallace Foundation.
- Leithwood, K. A., & Montgomery, D. J. (1982). The role of the elementary school principal in program improvement. *Review of Educational Research*, 52, 309–339.
- Lemmens, P., Knibbe, R., & Tan, R. (1988). Weekly recall and diary estimates of alcohol consumption in a general population survey. *Journal of Studies on Alcohol*, 53, 476–486.
- Lemmens, P., Tan, E., & Knibbe, R. (1992). Measuring quantity and frequency of drinking in a general population survey: A comparison of five indices. *Journal of Studies on Alcohol*, 49, 131–135.
- Louis, K. S., Marks, H., & Kruse, S. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal*, 33, 757–798.
- Mathison, S. (1988). Why triangulate? *Educational Researcher*, 17, 13–17.
- McLaughlin, M., & Talbert, J. E. (2006). *Building school-based TLCs: Professional strategies to improve student achievement*. New York, NY: Teachers College Press.
- Mintzberg, H. (1973). *The nature of managerial work*. New York, NY: Harper & Row.
- Mullens, J. E., & Gaylor, K. (1999). *Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction* (Working Paper No. 1999-08). Washington, DC: National Center for Education Statistics.

- Ogawa, R. T., & Bossert, S. T. (1995). Leadership as an organizational quality. *Educational Administration Quarterly*, 31, 224–243.
- Peterson, K. D. (1977). The principal's tasks. *Administrators Notebook*, 26(4), 1–4.
- Pitner, N. J. (1988). The study of administrator effects and effectiveness. In N. J. Boyan (Ed.), *Handbook of research in educational administration: A project of the American Educational Research Association* (pp. 99–122). New York, NY: Longman.
- Rosenholtz, S. J. (1989). *Teachers' workplace: The social organization of schools*. New York, NY: Longman.
- Ruddock, J., Chaplain, R., & Wallace, G. (1996). *School improvement: What can pupils tell us?* London, UK: Fulton.
- Scott, C. K., Ahadi, S., & Krug, S. E. (1990). *An experience sampling approach to the study of principal instructional leadership II: A comparison of activities and beliefs as bases for understanding effective school leadership*. Urbana, IL: National Center for School Leadership.
- Smithson, J., & Porter, A. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum—The reform up close study* (CPRE Research Report Series No. 31). Madison, WI: Consortium for Policy Research in Education.
- Spillane, J. (2006). *Distributed leadership*. San Francisco, CA: Jossey-Bass.
- Spillane, J., Camburn, E., & Pareja, A. (2007). Taking a distributed perspective to the school principal's work day. *Leadership and Policy in Schools*, 6, 103–125.
- Spillane, J., Halverson, R., & Diamond, J. (2001). Investigating school leadership practice: A distributed perspective. *Educational Researcher*, 30, 23–28.
- Spillane, J., Halverson, R., & Diamond, J. (2004). Towards a theory of school leadership practice: Implications of a distributed perspective. *Journal of Curriculum Studies*, 36, 3–34.
- Stone, A., Kessler, R., & Haythornthwaite, J. (1991). Measuring daily events and experiences: Decisions for researchers. *Journal of Personality*, 59, 575–607.
- Suchman, L. (1995). Making work visible. *Communications of the AMC*, 38(9), 33–35.
- Tannenbaum, R., Weschler, I. R., & Massarik, F. (1961). *Leadership and organization: A behavioral science approach*. New York, NY: McGraw-Hill.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.
- Tucker, R. C. (1981). *Politics as leadership*. Columbia, SC: University of Missouri Press.
- Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality*, 59, 339–354.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 10

Learning in Collaboration: Exploring Processes and Outcomes



Bénédicte Vanblaere and Geert Devos

10.1 Introduction

Given the major changes taking place in education over the past decades, professional development of teachers has become a necessity for teachers throughout their entire career (Richter, Kunter, Klusmann, Lüdtke, & Baumert, 2011). Historically, professional development activities of teachers were seen as attending planned and organized external professional development interventions, which generally assigned a passive role to teachers and was episodic, fragmented, and idiosyncratic (Hargreaves, 2000; Lieberman & Pointer Mace, 2008; Putnam & Borko, 2000). As such, these impediments and constraints limited the relevance of traditional professional development for real classroom practices (Kwakman, 2003).

Currently, many educational researchers argue that a key to strengthening teachers' ongoing growth and ultimately students' learning lies in creating professional learning communities (PLCs), where teachers share the responsibility for student learning, share practices, and engage in reflective enquiry (Slegers, den Brok, Verbiest, Moolenaar, & Daly, 2013). Hence, this represents a shift towards ongoing and career-long professional development embedded in everyday activities (Eraut, 2004), where learning is no longer a purely individual activity but becomes a shared endeavour between teachers (Lieberman & Pointer Mace, 2008; Stoll, Bolam, McMahon, Wallace, & Thomas, 2006). A significant body of research has attributed improvement gains, enhanced teacher capacity, and staff capacity at least in part to the formation of a PLC, thus demonstrating the relevance of teachers' collegial relations as a factor in school improvement (Bryk, Camburn, & Louis, 1999; Darling-Hammond, Chung Wei, Alethea, Richardson, & Orphanos, 2009; McLaughlin &

B. Vanblaere · G. Devos (✉)
Ghent University, Ghent, Belgium
e-mail: geert.devos@ugent.be

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_10

197

Talbert, 2001; Stoll et al., 2006; Tam, 2015; Vangrieken, Dochy, Raes, & Kyndt, 2015; Wang, 2015).

Previous studies on PLCs are rich in normative descriptions about what PLCs should look like (Vescio, Ross, & Adams, 2008). In reality, however, schools that function as strong PLCs and teachers that engage in profound collaboration with colleagues are few in number (Bolam et al., 2005; OECD, 2014). As such, it is not surprising that educationalists are keen to learn more about what characterizes schools in several developmental stages of PLCs and what teachers do differently in strong PLCs (Hipp, Huffman, Pankake, & Olivier, 2008; Vescio et al., 2008). Moreover, little is known about what teacher learning through collaboration in the everyday school context in PLCs looks like and which identifiable consequences collaboration can have for teachers' cognition and practices (Borko, 2004; Tam, 2015; Vescio et al., 2008). This leads to three different methodological challenges: First, it is necessary to identify schools in different developmental stages of PLCs. Second, it is important to have rich descriptions of how teacher learning through collaboration in schools takes place. This is a complex process that includes mental, emotional, and behavioural changes. This necessitates a long-term observation of the process. Third, it is important to compare this process in schools at different stages of PLC in order to identify what makes the difference between these stages. To address these complex challenges, we designed a mixed method study. In the first place, it was important to identify what categories of schools, related to the developmental stages of PLCs, can be distinguished using the three core interpersonal PLC characteristics. Next, we selected four cases from contrasting types of PLC schools. A year-long study was set up to contrast the collaboration and resulting learning outcomes of experienced teachers in two high and two low PLC schools. Few studies in the field of PLCs have adopted a mixed methods approach (Slegers et al., 2013), and studies about PLCs in primary education are lacking (Doppenberg, Bakx, & den Brok, 2012). This innovative mixed methods approach set in primary education wanted to explore, if the challenging methodological research goals were met and what the points of attention and pitfalls of this method were. In this respect, the study had both an empirical and a methodological aim.

10.1.1 PLC as a Context for Teacher Learning

In her seminal study about the conceptualization and measurement of the impact of professional development, Desimone (2009) argues that the core theory of action for professional development consists of four elements:

1. Teachers experience effective professional development.
2. This increases their knowledge and skills and/or changes their attitudes and beliefs.
3. Teachers use these new skills, knowledge, attitudes, and beliefs to improve the contents of their instruction or pedagogical approach.
4. These instructional changes foster increased student learning.

Many definitions of teacher learning and studies about the effects of professional development have confirmed that teacher change involves changes in cognition and in behaviour (Bakkenes, Vermunt, & Wubbels, 2010; Clarke & Hollingsworth, 2002; van Veen, Zwart, Meirink, & Verloop, 2010; Zwart, Wubbels, Bergen, & Bolhuis, 2009). Many professional development programs follow an implicit causal chain and assume that significant changes in practice are likely to take place only after mental changes are present. However, this idea has been criticized and contested for quite some time by authors pointing out that a mental change does not necessarily have to result in a change of behaviour to be seen as learning, nor does a change in behaviour have to lead to mental changes (Meirink, Meijer, & Verloop, 2007; Zwart et al., 2009). As such, more interconnected models that adopt a cyclic or reciprocal approach have been presented (Clarke & Hollingsworth, 2002; Desimone, 2009).

As for teacher behaviour as a learning outcome, teacher learning is strongly connected to professional goals that stimulate teachers to continuously seek improvement of their teaching practices (Kwakman, 2003). In this study, changes in teacher behaviour are thus described in terms of changes in teachers' classroom teaching practices (e.g. changed contents of instruction, or changes in pedagogical approach). According to Bakkenes et al. (2010), it is important to also take into account teachers' intentions for practices as learning outcomes, as these can be seen as precursors of change in actual practice. Regarding the mental aspect of learning outcomes, learning opportunities are expected to result in changes in teacher competence, seen as a complex combination of beliefs, knowledge, and attitudes (Deakin Crick, 2008; van Veen et al., 2010). For instance, Bakkenes et al. (2010) identified changes in knowledge and beliefs (new ideas and insights, confirmed ideas, awareness) and changes in emotions (negative emotions, positive emotions) in their research.

Studies acknowledge the difficulty of change, both in cognition and in behaviour (Bakkenes et al., 2010; McLaughlin & Talbert, 2001; Tam, 2015). Nevertheless, PLCs hold particular potential in this regard as documented by studies that link these collaborative learning opportunities to teacher change (Bakkenes et al., 2010; Hoekstra, Brekelmans, Beijaard, & Korthagen, 2009; Tam, 2015; Vescio et al., 2008). However, few authors focus on learning outcomes related to both cognition and behaviour in the same study.

Although a universally accepted definition of PLCs is lacking (Bolam et al., 2005; Stoll et al., 2006; Vescio et al., 2008), a common denominator can be identified: Collaborative work cultures are developed in PLCs, in which systematic collaboration, supportive interactions, and sharing of practices between stakeholders are frequent. These communities strive to stimulate teacher learning, with the ultimate goal of improving teaching to enhance student learning and school development (Bolam et al., 2005; Hord, 1997; Louis, Dretzke, & Wahlstrom, 2010; Slegers et al., 2013; Vandenberghe & Kelchtermans, 2002).

Parallel to the diversity in definitions, studies about PLCs differ greatly with regard to the operationalization of the concept. However, several often-cited features of PLCs can be found, related to what Slegers et al. (2013) identified as the interpersonal capacity of teachers. This interpersonal capacity encompasses

cognitive and behavioural facets. Related to the cognitive dimension, many scholars point to a collective feeling of responsibility for student learning in PLCs (Bryk et al., 1999; Hord, 1997; Newmann, Marks, Louis, Kruse, & Gamoran, 1996; Stoll et al., 2006; Wahlstrom & Louis, 2008). Concerning the behavioural dimension, strong PLCs are characterized by reflective dialogues or in-depth consultations about educational matters, on the one hand, and deprivatized practice, on the other hand, through which teachers make their teaching public and share practices (Bryk et al., 1999; Hord, 1997; Louis & Marks, 1998; Stoll et al., 2006; Visscher & Witziers, 2004). Time and space are provided in successful PLCs for formal collaboration (i.e. collaboration that is regulated by administrators, often compulsory, implementation-oriented, fixed in time, and predictable) as well as informal collaboration (i.e. spontaneous, voluntary, and development-oriented interactions) (Hargreaves, 1994; Stoll et al., 2006). However, due to the conceptual fog surrounding the operationalization of the concept, empirical evidence documenting these essential PLC characteristics is lacking (Vescio et al., 2008).

While the idea behind PLCs receives broad support and many principals make strong efforts to promote collegial cultures in their schools, the TALIS 2013 study (OECD, 2014) showed that teachers still work in isolation from their colleagues for most of the time. Opportunities for developing practice based on discussions, examinations of practice, or observing each other's practices remain limited. Teachers tend to share practices (Meirink, Imants, Meijer, & Verloop, 2010), but often through conversations that stay at the level of planning or talking about teaching (Kwakman, 2003) or through collaboration that lacks profound feedback among teachers (Svanbjörnsdóttir, Macdonald, & Frímannsson, 2016). Others have found that collaboration is often confined to solving problems that arise in the day-to-day practice (Scribner, 1999), while it is crucial in strong PLCs to also exchange and discuss teachers' personal beliefs (Clement & Vandenberghe, 2000). It is necessary to distinguish between different forms and levels of collaboration as the benefits associated with it are not automatically achieved by any type of collaboration (Little, 1990). Studies highlight that collaboration between teachers should meet some standards in order to lead to profound teacher learning (Meirink et al., 2010). This is exemplified by the work of Hord (1986), who distinguished between two types of collaboration. On the one hand, she defined collaboration as actions in which two or more teachers agree to work together to make their private practices more successful but maintain autonomous and separate practices. On the other hand, teachers can work together while being involved in shared responsibility and authority for decision-making about common practices. These types are related to, respectively, the efficiency dimension of learning, where teachers mainly achieve greater abilities to perform certain tasks, and the innovative dimension, which results in innovative learning and requires the replacement of old routines and beliefs (Hammerness et al., 2005). While the former type of learning and collaboration is found in almost all schools, it is the latter type that characterizes practices in PLCs. As such, it is important to identify how collaboration in schools in diverse PLC development stages manifests. Studies that closely monitor interactions between teachers in primary education are lacking (Doppenberg et al., 2012).

10.1.2 The Study (Mixed Methods Design)

The above literature shows that our knowledge is still limited about the way a PLC can contribute to experienced primary school teachers' changes in cognition and behaviour. A mixed methods research design is adopted in this study, in which we combine both qualitative and quantitative methods into a single study (Leech & Onwuegbuzie, 2009). This study is based on an explanatory sequential design (Greene, Caracilli, & Graham, 1989). We opted for this mixed methods design because of the different methodological challenges we faced. First, we wanted to identify different developmental stages of PLCs, in which primary schools can be situated (RQ1). For this challenge, we needed a substantial set of primary schools, in which quantitative data were collected. This quantitative method in a large sample of schools was necessary to identify different categories of PLCs based on the three interpersonal PLC characteristics: Collective responsibility, deprivatized practice, and reflective dialogue (Wahlstrom & Louis, 2008). A survey among the teaching staff of these schools provided the data for these characteristics. The aggregation of the data for each school enabled us to identify four meaningful and useful clusters that reflect different developmental stages of PLCs.

A second methodological challenge is to provide rich descriptions of teacher learning through collaboration on a long-term basis and to understand how this differs between different developmental stages of PLCs. To meet this challenge, the method of following-up on outliers or extreme cases is then used in the qualitative part of this study (Creswell, 2008). We compare the type and contents of the year-long collaboration of experienced teachers about a school-specific innovation in four schools in extreme clusters (high presence versus low presence of PLC characteristics; RQ2). We also compare how teachers in these four schools look back at the collaboration and how they assess the quality of the collaborative activities (RQ3). Furthermore, we investigate how PLCs can contribute to experienced teachers' learning (RQ4), more particularly to cognitive and behavioural changes, thus deepening the general framework of learning outcomes of Bakkenes et al. (2010). We focus on experienced teachers as this allows us to gain insight into learning outcomes that go beyond merely mastering the basics of teaching (Richter et al., 2011). Using a longitudinal perspective through digital logs enables us to focus on differences between high and low PLC schools in the evolution of collaboration and learning outcomes throughout one school year. The choice of using digital logs as a qualitative method was inspired by the study of Bakkenes et al. (2010). In this study, digital logs were used to ask teachers to describe learning experiences over a period of one year. This procedure displayed several strengths: The provision of rich descriptions of teacher learning that enabled the researchers to differentiate between (different) experiences of teachers, an efficient way of collecting qualitative data with the same time-intervals from a relative large number of participants, the opportunity to collect similar information (similarly structured with different time-intervals) and comparable data across different schools, and the opportunity to collect longitudinal data over a one-year period.

The methods and results for the quantitative and qualitative research phase are discussed separately. The findings are interpreted jointly in the discussion.

10.2 Quantitative Phase

10.2.1 Methods

An online survey was completed by 714 Flemish (Belgian) primary school teachers from 48 schools. On average, 15 teachers per school completed the questionnaire, with a minimum of 3 teachers in each school. The mean school size was 21 teachers (range: 6–42 teachers) and 298 students (range: 100–582 students). As for the teachers, the sample included 86% female teachers, which is similar to the male-female division in Flemish primary schools. Teachers' experience in the current school ranged from 1 to 38 years ($M = 13$ years), while the experience in education varied from 1 to 41 years ($M = 16$ years).

To measure the interpersonal PLC characteristics (Slegers et al., 2013), we used three subscales of the 'Professional Community Index' (Wahlstrom & Louis, 2008): collective responsibility, deprivatized practice, and reflective dialogue (Vanblaere & Devos, 2016). A summary of the main characteristics of the scales can be found in Table 10.1.

As a first step in the analysis, aggregated mean scores for the three PLC characteristics were computed. The intraclass correlations of a one-way analysis of variance with a cut-off score of .60 (Shrout & Fleiss, 1979) were used to determine that it was legitimate to speak of school characteristics (see ICC in Table 10.1). Then, a two-step clustering procedure was performed with SPSS22 to attain stable and interpretable clusters that have maximum interpretable discrimination between the different clusters (Gore, 2000). First, the three aggregated PLC characteristics were standardized and entered in a hierarchical cluster analysis, using Ward's method on squared Euclidean distances, which minimizes within-cluster variance. Second, the

Table 10.1 Summary of the scales

	N_{items}	$M(SD)$	α	ICC	Example item	Range
Collective responsibility	3	3.68(.66)	.68	.83	Teachers in this school feel responsible to help each other improve their instruction.	Strongly disagree (1) – Strongly agree (5)
Deprivatized practice	3	1.91(.75)	.74	.75	How often in this school year have you had colleagues observe your classroom?	Never (1) – Very often (5)
Reflective dialogue	5	3.26(.70)	.76	.72	How often, in this school year, have you had conversations with colleagues about the development of a new curriculum?	Never (1) – Very often (5)

cluster centres from the hierarchical cluster analysis were used as non-random starting points in an iterative k-means (non-hierarchical) clustering procedure. This process permitted the identification of relatively homogeneous and highly interpretable groups of schools in the sample, taking the three PLC characteristics into account.

10.2.2 Results

In the first step of the cluster analysis, the cluster division had to explain a sufficient amount of the variance in the three PLC characteristics. We estimated cluster solutions with two to four clusters and inspected the percentage of explained variance in each solution (Eta squared). As only the four-cluster solution explained more than 50% of the variance in all three variables, the other cluster solutions were not considered further. Step two of the process was applied to the four-cluster solution, which yielded four clearly distinct clusters with sufficient explained variance (collective responsibility (.68), deprivatized practice (.63), and reflective dialogue (.77)). Table 10.2 presents a detailed description of these clusters, including standardized means, standard deviations, and descriptions.

Cluster 1 consisted of only 4 schools (8.4%) of the research sample. These schools reported high scores in all three interpersonal PLC characteristics, including deprivatized practice. This separates them from the schools in cluster 2 (n = 11, 22.9%), in which the scores were high for collective responsibility and reflective dialogue, but only average for deprivatized practice. This implies that teachers rarely observe each other's practices in cluster 2, while this occurs every now and then in the first cluster. Cluster 3 consisted of 22 schools (45.8%) scoring rather average on all three PLC characteristics. In these schools, teachers feel more or less collectively responsible for their students, engage in reflective dialogue every now and then, but rarely observe each other's teaching practice. Cluster 4 was also represented by 11 schools (22.9%) and showed a low presence of PLC characteristics.

Table 10.2 Standardized mean scores and standard deviations

	Cluster 1 (n = 4)	Cluster 2 (n = 11)	Cluster 3 (n = 22)	Cluster 4 (n = 11)
Collective responsibility	1.38 (.36)	.83 (.52)	-.06 (.50)	-1.22 (.82)
+		+	0	-
Deprivatized practice	2.33 (.49)	.20 (.47)	-.12 (.72)	-.79 (.61)
+		0	0	-
Reflective dialogue	1.12 (.60)	1.08 (.38)	-.10 (.52)	-1.29 (.48)
+		+	0	-
Cluster names	<i>High presence of all PLC characteristics</i>	<i>Average deprivatized practice; high collective responsibility and reflective dialogue</i>	<i>Average presence of all PLC characteristics</i>	<i>Low presence of all PLC characteristics</i>

10.3 Qualitative Phase

10.3.1 Case Selection and Method

In this part of the study, a multiple case study design was adopted. A purposeful sampling of extreme cases was carried out (Miles & Huberman, 1994), involving schools from cluster 1 with a strong presence of all PLC characteristics (high PLC) and schools from cluster 4 with a low presence of all PLC characteristics (low PLC). These schools were contacted, and we inquired about plans to implement an innovation or change during the following school year with implications for teachers' ideas, beliefs, and teaching practices. The final sample consists of four schools (two of high PLC and two of low PLC) that met this criterion and where teachers agreed to participate in the study.

The sample consists of 29 experienced teachers with at least five years of experience in education and three years of experience in the current school, based on Huberman's (1989) classification. The only exception is school D, where a teacher with only two years of experience in the current school also participated, since this teacher played a central role in the ongoing innovation. In school A, B, and D, all experienced teachers took part in the study. In school C, however, six of the experienced teachers involved in the innovation were randomly selected by the principal. Table 10.3 presents some context information on the four selected schools.

Teachers in the participating schools were asked to complete digital logs at four time-points over the course of one school year, i.e. at the beginning of the school year and at the end of each of the three trimesters (December, April, and June). In total, we received 109 completed logs (response rates $\geq 90\%$, see Table 10.3). The first log was intended to provide the authors with more background information about the antecedents, implementation, and consequences of the innovation. The focus of this study was on the remaining three logs ($n = 80$), in which teachers were asked about their collaborative activities concerning the innovation during that trimester and the resulting learning outcomes. More specifically, teachers were first asked to list the different kinds of collaborative activities they had actively engaged in and to describe the nature and contents of these activities. Teachers had the option to fill in any type of activity while being provided with some examples (e.g. discussing the innovation at a staff meeting, jointly preparing and evaluating a lesson with regards to innovation, informal discussion with colleagues during break-time). They were also instructed to list activities separately, if the stakeholders differed. Teachers could list from one to ten different kinds of activities. For each activity they undertook, the teachers received brief, structured follow-up questions about the collaboration process. Each question had to be answered separately, prompting the teachers to provide additional information about the stakeholders in the described collaborative activity, who initiated it, where and when it took place, how frequently it occurred, and any constraints they experienced. Secondly, teachers were asked in each log to reflect upon what they had learned through this collaboration and to describe the contribution to their own classroom practices and their competence as

Table 10.3 Background information on the case study schools

	Cluster 1 High PLC		Cluster 4 Low PLC	
	HIGH A	HIGH B	LOW C	LOW D
<i>General school characteristics:</i>				
Alternative school	Yes, Freinet	No	No	No
Total number of teachers	8	15	25	8
Number of students	100	240	376	130
School population	High SES students	Moderately high SES students	Moderately high SES students	High SES students
Innovation	New teaching method (language)	New teaching method (technique)	New teaching method (language: reading)	Incorporation of cross-curricular 'learning to learn' in all subjects
<i>Digital logs' respondents (experienced teachers):</i>				
Number of participating teachers	4 female, 0 male	11 female, 1 male	5 female, 1 male	6 female, 1 male
Average years of experience in education	16	18	22	15
Average years of experience in current school	14	15	19	12
Response rate	94%	98%	92%	90%

a teacher. This was an open question, but teachers were nonetheless instructed to mention how each collaborative activity had contributed to these outcomes. Responses to this question varied from 10 to 394 words. In the final log, all teachers were asked to briefly discuss their general appreciation of the quality of their own collaboration over the past year. Responses to this question varied from two-worded expressions (e.g. 'Great collaboration!') to 233 words.

The logs were coded using within- and cross-case analysis (Miles & Huberman, 1994). The first round of data analysis examined each separate log, which was treated as a single case. Considerable time was spent on the process of reading and re-reading the logs, as they were submitted throughout the year, in order to assess the meaningfulness of the constructs, categories, and codes (Patton, 1990). If the log of a teacher was unclear, contributions of other teachers at the school were searched through for possible clarifications. Additional information from teachers was requested by e-mail or telephone, when needed, to ensure a correct interpretation.

A coding scheme was developed based on the theoretical framework and based on themes emerging from the data itself. The categories used to identify features of collaboration were: (1) type (discussions about practice, teaching together or sharing teaching practices, working on teaching materials, practical collaboration, and

no collaboration), (2) structure (formal and informal), (3) stakeholders (the entire school team, a fixed sub-team, interactions between two or three teachers, and external stakeholders), and (4) duration (frequency and recurrence throughout the year). The reflections of the teachers on the collaboration at the end of the year were divided into positive or negative impressions based on indicators of appreciation in the language used. The coding framework used to categorize the outcomes of the collaboration: No learning outcome, changes in knowledge and beliefs (new ideas and insights, confirmed ideas, awareness), changes in practices (new practices, intentions for new practices, alignment), changes in emotions (negative emotions, positive emotions), and general impression of contribution. Each log was assessed with regard to the presence of these outcomes. Related to the coding of 'new practices,' it should be noted that logs were only coded as containing new practices when these changes were a consequence of the collaboration between teachers. Nevertheless, certain collaborative activities, in essence, also implied new classroom practices, even though they were not coded as such (e.g. co-teaching with coaches (HIGH B) and lesson observation and workshops (LOW D)). A second researcher, who was not familiar with the study or participating schools, was trained to grasp the meaning of the coding and coded 30% of the logs ($n = 24$). The intercoder-reliability was .89, which is in accordance with the standard of .80 of Miles and Huberman (1994).

Once all separate logs were coded, data from teachers within the same school were combined to provide an overview of the collaboration and learning outcomes at each school in the first, second, and third trimester. Similarly, teachers' general appreciation of the quality of their own collaboration, as written down in the final log, was described for each participating school. This resulted in a school-specific report that summarized all findings for each school. As a member check the school-specific report was sent to the principal, accompanied by the request for discussing this report with their teachers and to provide us with feedback. This allowed principals and teachers to affirm that these summaries reflected the processes that occurred throughout the school year at their school. No alterations were requested, thus confirming the completeness and accuracy of the study. Next, the within-case analysis was extended by comparing the logs over time for each school. Fourth, a cross-case analysis was conducted, where the four schools were systematically compared with each other to generate overall findings that transcend individual cases and to identify similarities and differences between high and low PLC schools; Nvivo10 was used to organize our analysis.

10.3.2 Results

10.3.2.1 Collaboration Between Teachers

Our results indicate that collaboration was shaped in a very different way in the two schools selected from the cluster with a high presence of PLC characteristics (high PLC) and in the two schools from the cluster showing a low presence of PLC characteristics (low PLC). In the following paragraphs, the differences in the type of collaborative activities will be explained more in depth, with an explicit focus on the evolution of practices throughout the school year.

A first major difference between the high and low PLC schools lies in teachers making their teaching public by engaging in deprivatized practice, or working on teaching materials together in high PLC schools. However, the execution of these shared practices differed between both high schools. In HIGH B, several teachers were appointed as coaches, specifically for the implementation of the innovation. Each coach was paired with one or two teachers from adjacent grades, and they engaged in several structured cycles of collaboration. In the first and second trimester, coaches and teachers worked on lesson preparations together or in consultation, by frequently discussing the design, contents, and pedagogical approach of the lessons that were taught related to the innovation. These lessons were then taught through co-teaching or taught by one teacher and observed by the other. At the initiative of several teachers using the innovation in their daily practice, a sub-team of teachers in HIGH A developed classroom materials together throughout the school year. In addition, HIGH A was visited in the third trimester by a teacher from a school working with the same innovation as well as by a group of teachers interested in implementing the innovation in the future. Artefacts, classroom practices, information, and findings about the implementation of the innovation were shared with these external stakeholders. As such, these practices illustrate that deprivatized practice can occur both within schools and between schools. This is in contrast with the low PLC schools, where such practices were virtually non-existent, apart from a one-time lesson observation in LOW D between two teachers, with no real follow-up.

A second difference relates to practical collaboration between teachers. In the low PLC schools, it was common for teachers to engage in basic practical collaboration. This was especially the case throughout the school year in LOW C, where teachers from the same grade, for instance, visited the library together or assessed students' reading level together with the special needs teacher. Remarkably, this is even the only type of collaboration that multiple teachers of LOW C mentioned in the third trimester of the school year. Several teachers in LOW D mainly had practical interactions at specific moments (e.g. at the end of the school year), or with external stakeholders (e.g. a volunteer, who taught weekly chess lessons in two classrooms).

Third, our results show that while teachers in both high and low PLC schools participated in discussions about how to incorporate the innovation in their daily practice, the extent of these conversations differed noticeably. Teachers in all

schools described dialogues with specific partners (i.e. teachers of the same grade, adjacent grades, or coach) about general and practical matters. In low PLC schools, most interactions were limited to these fixed partnerships, and discussions about the innovation with the entire team at staff meetings were mentioned infrequently in the logs of teachers, indicating a low ascribed importance of these meetings. Structured sub-teams of teachers were largely absent in low PLC schools, with the exception of two working groups in LOW D. These working groups were launched at the end of the school year, met once, and were focused on practical arrangements and requests of teachers for the following school year. In contrast, in high PLC schools, conversations about day-to-day problems or questions involving the innovation were also frequently discussed spontaneously with colleagues in between lessons (or at lunch-time) with whoever was present. Teachers also systematically brought up that the innovation was discussed during staff meetings throughout the school year. Both high PLC schools had a structured sub-team of teachers (coaches in HIGH B, teachers using the innovation daily in HIGH A). Additionally, teachers in these schools exchanged experiences and expertise with teachers from other schools implementing a similar innovation and receiving external assistance, either on a structural regular basis (HIGH B) or in a one-time workshop (HIGH A).

Furthermore, most dialogues occurred in the low PLC schools in the first trimester, after which the frequency of conversations about the innovation diminished drastically. Contrarily, dialogues in high PLC schools were maintained across the school year.

The contents of dialogues usually remained at a superficial level in low PLC schools, as illustrated by teachers in LOW C, who stated that initial staff meetings were about making arrangements and expressing expectations regarding the innovation, while this evolved throughout the school year into reminders for teachers to implement the innovation.

However, teachers in the high PLC schools did engage in several kinds of profound and reflective dialogues. For instance, each coach in HIGH B completed a structured evaluation with their partner each time they had jointly prepared and taught a lesson. At the end of the school year, they reflected upon the implementation of the innovation and the link between the innovation and other teaching contents. Additionally, both sub-teams of teachers in the high PLC schools had several formal meetings each trimester as well as informal discussions during breaks or outside of school hours, aimed at monitoring and moving the innovation forward. Furthermore, staff meetings with the entire team were used as a way to facilitate planning, but most importantly to share teachers' beliefs, opinions, and experiences.

In conclusion, the results show several substantial differences between the high and low PLC schools in their collaboration. While teachers in all schools engaged in day-to-day conversations about the implementation of the innovation, these dialogues were more sustained throughout the school year and more spread throughout the entire team in high PLC schools. Additional collaboration was also of a higher importance in high PLC schools compared to low PLC schools, involving activities, such as deprivatized practice, discussions with the entire team, developing teaching materials, and having profound conversations about beliefs and experiences. High

PLC schools also undertook meaningful partnerships with external stakeholders, while low PLC schools regularly engaged in practical collaborations. With regard to the initiators of collaboration, high PLC schools appear to make good use of both structured formal collaboration and spontaneous informal collaboration, while the initiative of collaboration often remained with individual teachers in low PLC schools.

10.3.2.2 Learning Outcomes from the Collaboration

With regard to the final qualitative research question, teachers mentioned a wide range of outcomes when asked what they had learned through interacting with their colleagues. In total, ten different types of outcomes were distinguished in teachers’ logs. Table 10.4 provides an overview of the occurrence of the outcomes throughout the school year. The communalities and differences between the contents and the diversity of learning outcomes in high and low PLC schools are discussed and illustrated in the following paragraphs.

Content of the Outcomes

We first describe the outcomes that are marked as frequently mentioned in Table 10.4 (i.e. general impression of contribution, no outcome, new ideas, new practices, and changes in alignment), after which we move on to a brief discussion of the

Table 10.4 Learning outcomes per school throughout the school year

	High A		High B				Low C			Low D		
	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3
General impression of contribution	***	*	***	***	***	*	**	*	***	***	**	
Negative emotions						*						
Positive emotions	*	*		**	*	*						
New ideas	**	*	**	**	**	***	***	*	*	***	*	
Confirmed ideas					*			*	*			
Awareness				**	**					*		*
New practices	**	*	**	**	**	*	***	***	*	*		
Intentions new practices			*	*		**						**
Changes in alignment	**	*	**			**				**	*	***
No outcome			*		*	*		***	***	*	***	*

Note: T1 = trimester 1, T2 = trimester 2, T3 = trimester 3
 ***represents the most frequently mentioned outcome during that trimester (in case of a tie, two outcomes are indicated);
 **represents outcomes mentioned by multiple teachers during that trimester,
 *represents outcomes mentioned by one teacher during that trimester

remaining outcomes (i.e. positive emotions, intentions for practices, awareness, negative emotions, and confirmed ideas).

Teachers from both high and low PLC schools mentioned that their collaboration somehow contributed to their professional growth. This positive impression is most consistent throughout the school year in the high PLC schools. However, not all teachers had the impression that the collaboration made meaningful contributions to their competence or practices, especially in low PLC schools. Logs from the second and third trimester in these low PLC schools show a lack of learning outcomes stemming from collaboration for a considerable group of teachers. Several teachers merely explained their collaborative activities again or mentioned what students had learned, but failed to provide evidence of their own learning outcomes.

Our results indicate that new ideas, insights, and tips as a learning outcome occur consistently in high and the low PLC schools throughout the school year, as only the logs of the third trimester in LOW D did not contain any new ideas. Here, we did not find any systematic differences between high and low PLC schools.

New practices, as a result of collaboration, were mentioned several times in the high PLC schools. In the low PLC schools, no profound changes were reported. New practices at a basic level were the most frequently mentioned outcome for LOW C in the first two trimesters, usually as a result of practical collaboration, which was strongly present at this school. Teachers in LOW D hardly mentioned new practices of any nature.

Furthermore, our results suggest differences between schools regarding the stakeholders in aligning practices between teachers. This type of outcome transcends the individual classroom practice of teachers and refers to classroom practices being geared to one another. However, these results should be interpreted with caution as changes in alignment occurred systematically in two schools only (HIGH A, and LOW D). In the high PLC school, teachers spoke of aligning practices for the whole school during the school year, for example: "It was a useful meeting to exchange experiences and to find common ground. Practices were geared to one another." (Teacher, HIGH A). In LOW D, this practice was not spread throughout the school as most of the statements could be attributed to two teachers, who consistently mentioned aligning practices throughout the year. One teacher explained: "I got a clear image of what the testing period in grades 4 and 6 looks like. This allowed us to discuss the learning curve we want to implement: increasing difficulty level, what is expected in the next year,...." Only at the end of the school year, teachers mentioned aligning practices for the entire school in a one-off working group.

Although not mentioned frequently, it is noteworthy that positive emotions were only reported in the high PLC schools. Several teachers expressed throughout the year that they felt supported by their colleagues, coaches, or principal, and that they were glad that help from colleagues was available.

Finally, our results show that collaborative interactions between teachers only rarely lead to negative emotions (e.g. feelings of concern and doubt about the role as coach for the following years) or confirmed ideas, in both high and low PLC schools.

Diversity of the Outcomes

Looking at the diversity of reported outcomes in schools (see Table 10.4), teachers in the high PLC schools, on average, mentioned multiple of the outcomes described above as a result of collaboration during each trimester. Hence, teachers from high PLC schools have, in general, attained more varied learning outcomes per trimester than teachers in low PLC schools. Over the three trimesters, teachers in HIGH A, and HIGH B consistently mentioned multiple outcomes per trimester and thus combinations of learning outcomes. In HIGH B, the full range of outcomes was reached, as every outcome was mentioned by at least one teacher at some point in time during the school year.

However, outcomes were less diverse in low PLC schools. In general, these teachers did not describe any changes in their competence or practices, or indicated just one outcome (e.g. new practices, new ideas). This trend was present throughout the year in LOW C, while outcomes were more diverse in the first trimester in LOW D, but then diminished drastically in the second and third trimester.

10.4 Discussion and Conclusion

Combining quantitative and qualitative data in this study, allowed us to ‘dig deeper’ into the question of how PLCs function and contribute to teachers’ learning outcomes, resulting in generalizable findings as well as detailed and in-depth descriptions of key mechanisms in several schools that were followed throughout an entire school year. In particular, we quantitatively examined, which types of primary schools can be distinguished, based on the strength of three interpersonal PLC characteristics. This resulted in four meaningful categories of PLCs at different developmental stages. Subsequently, we qualitatively documented the collaboration and resulting learning outcomes of experienced teachers related to a school-specific innovation over the course of one school year at four schools at both ends of the spectrum (high PLC versus low PLC). Our analyses showed the following key findings:

The first research question was aimed at analysing into which categories primary schools could be classified based on the strength of three interpersonal PLC characteristics (collective responsibility, reflective dialogue, and deprivatized practice). Cluster analysis revealed four meaningful categories, reflecting different developmental stages: High presence of all characteristics (8.4% of schools); high reflective dialogue and collective responsibility, but average deprivatized practice (22.9%); average presence of all characteristics (45.8%); and low presence of all characteristics (22.9%). This confirms that there are considerable differences between schools in the extent to which they function as a PLC, with most schools in the stage of developing a PLC (Bolam et al., 2005). This classification is in line with previous categories found for Math departments in Dutch secondary schools that also

identified a high PLC cluster, a low PLC cluster, a deprivatized practice cluster, and an average cluster (Lomos, Hofman, & Bosker, 2011).

With our second research question, we wanted to clarify what characteristics of collaboration differed throughout the school year in schools with a high and low presence of all PLC characteristics, when dealing with a school-specific innovation. In this regard, our results confirmed previous studies that point to the frequent occurrence of basic day-to-day discussions about problems and teaching (Meirink et al., 2010; Scribner, 1999). However, based on our knowledge, this study is one of the first ones to pinpoint differences between the high and low PLC schools in these lower levels of collaboration, such as storytelling and aid (Little, 1990). We add to the literature by concluding that teachers in low PLC schools talk about an innovation mainly at the start of the school year, albeit with varying frequencies. The occurrence of these dialogues strongly diminished throughout the school year at low PLC schools, while they were more common and sustained at the high PLC schools. In some cases, the contents of the dialogues can explain, why conversations were mostly limited to the first trimester (e.g. conversations about “students’ transition between grades, fieldtrips, planning of the year or tests, and communal year themes” in LOW D). Furthermore, dialogues at the low PLC schools occurred mostly with a fixed partner, whereas spontaneous conversations spread throughout the team were equally found at the high PLC schools. Hence, this suggests that characteristics that are mainly associated with higher order collaboration in successful PLCs (e.g. spontaneous and pervasive across time (Hargreaves, 1994)), are also present in ongoing basic interactions in high PLC schools. Additionally, only teachers at the low PLC schools mentioned practical collaboration with colleagues, for example, visiting a library together.

In contrast, collaboration at the high PLC schools went well beyond these day-to-day conversations or practical collaboration, as we expected based on research of, for instance, Bryk et al. (1999), Little (1990), and Bolam et al. (2005). In this regard, our study shows that deprivatized practice can occur with a variety of stakeholders, as teachers opened up their classroom doors and made their teaching public, either for teachers from their own school (HIGH B) or teachers from other schools (HIGH A). In relation to the latter, it is remarkable that both high PLC schools were strong in building partnerships with other schools and sharing their experiences as well as making use of external support. This is in line with the idea that external partnerships can help a PLC to flourish (Stoll et al., 2006). Teachers were also responsible for developing concrete materials, such as lesson plans, that could be used by the team, which increases the level of interdependence in the team according to Meirink et al. (2010).

Furthermore, spontaneous as well as regulated reflective dialogues in small groups occurred. These included in-depth spontaneous reflections with an intention of improving practices throughout the entire school. Moreover, the importance of staff meetings and sub-teams as collaborative settings (Doppenberg et al., 2012) was confirmed for the high PLC schools. In particular, staff meetings were much more meaningful at the high PLC schools compared to low PLC schools, as meetings took place throughout the school year and left room for discussing teachers’

beliefs, experiences, and suggestions. Clement and Vandenberghe (2000) and Achinstein (2002) previously pointed to the importance of discussing beliefs for continual growth and renewal in schools. A possible explanation for the finding that collaboration often does not go beyond practical problem-solving and avoids discussions about beliefs at low PLC schools can be found in the field of micro-politics. Collaboration that includes talk about values and deeply held beliefs, requires a safe environment of trust and respect, but also increases the risk of conflict and differences in opinion (Johnson, 2003). According to Achinstein (2002), it is important to balance maintaining strong personal ties, on the one hand, while sustaining a certain level of controversy and differences in opinion, on the other hand.

It is interesting that both high PLC schools proactively installed a structured sub-team of teachers, intended to steer and monitor the innovation. Regardless of whether such a team is put together for the innovation (HIGH B), or existed previously (HIGH A), we think that this contributed greatly to the overall quality and continuation of collaboration at these schools, as interactions were not merely left to the initiative of individual teachers. This complements the finding of Bakkenes et al. (2010) and Doppenberg et al. (2012), who suggested that organized learning environments are qualitatively better than informal environments.

The third research question covered differences in teachers' appreciation of the general quality of their own collaboration. Remarkably, almost all teachers expressed a positive feeling about the collaboration, even in low PLC schools. This leads to an important methodological suggestion, namely that caution is required when dealing with teachers' perceptions of the quality of collaboration as in indicator of actual collaboration, because this can be an over-estimation of reality. A more accurate picture can be obtained, for example, by inquiring about the type and frequency of collaboration.

The final research question dealt with the differences in learning outcomes between the high and low PLC schools. The most striking difference is located in the diversity of outcomes that teachers reported. - More specifically, learning outcomes were overall more diverse and numerous throughout the school year for the high PLC schools compared to the low PLC schools. The sharp drop in learning outcomes in one of the low PLC schools in the second trimester might be due to the decrease of dialogues throughout the year in the low PLC schools. In relation to the contents of the learning outcomes, our results add to the general learning outcomes framework of Bakkenes et al. (2010) by expanding it to learning outcomes resulting solely from collaboration and exploring the occurrence of the outcomes at high and low PLC schools. Unsurprisingly, not all collaboration resulted in learning outcomes, especially at the low PLC schools. However, the logs showed that both at the high and low PLC schools, collaboration frequently led to new ideas and insights, or a general impression that the collaboration had made a contribution. This is in line with the finding of Doppenberg et al. (2012), who noted that teachers often mention implicit or general learning outcomes. A possible explanation for this is that both outcomes are fairly easy to achieve and non-committal towards the future. Another possibility is that teachers mainly associate learning with changes in cognition or the general impression of having learned something; it is also imaginable

that it was difficult for teachers to express what they had learned exactly, leading them to report a general impression. Nevertheless, new practices in line with the ongoing innovation also emerged. At the low PLC schools, new practices were limited, or mainly identified, as practical changes in classroom practices, or what Hammerness et al. (2005) referred to as 'the efficiency dimension of learning.' Only the collaboration at the high PLC schools seemed powerful enough to also provoke profound changes in practices or the innovative dimension of teacher learning (Hammerness et al., 2005). Additional intentions for practices were mainly identified at the end of the school year. Changes in emotions, confirmed ideas, changes in alignment, and awareness occurred rarely as learning outcomes. In conclusion, our results confirm that collaboration can result in powerful and diverse learning outcomes (Borko, 2004), but that this is not an automatic process for all collaboration (Little, 1990).

As with all research, there are some limitations to this study that cause us to be prudent about our findings. First, an explanatory sequential mixed methods design was used in this study. As such, our case studies were purposefully sampled based on available quantitative data. While this has many advantages, it implied that we had certain expectations regarding the collaboration in these schools beforehand, influencing our interpretation of the qualitative results. As such, we believe in the value of several precautions to limit this possible bias, as explained in the methods section (e.g. member check, the use of double-coding).

Second, the qualitative results are based on digital logs completed by teachers throughout the year. Individual perceptions were combined with the logs of other teachers from the school, when possible (e.g. for collaboration), and individual listings were seen as an indicator of the ascribed relevance of activities, but our study nevertheless relied heavily on self-report. Furthermore, some teachers did not provide detailed information about the nature of changes in practices or cognition resulting from the collaboration, especially at low PLC schools. As the logs were more elaborate at high PLC schools, this might have influenced our findings. In this regard, future research could add useful information by combining digital logs with interviews, or observations of collaboration and resulting changes to obtain more similar information from all teachers. Moreover, this study generally refrains from linking specific collaboration to certain outcomes, because not all teachers described their learning outcomes separately for each collaborative activity. Bearing in mind that it can be difficult for teachers to pinpoint what they have learned exactly, future research could address this gap.

Third, the case studies offer insight into experienced teachers' collaboration and learning at four primary schools that were selected through extreme case sampling and have rather unique profiles. Furthermore, the high average in years of teaching experience at the school, combined with the fairly small school sizes, point to rather long-term relationships between the participating teachers, which likely played a role in our results. Additionally, some collaboration with beginning teachers was mentioned by experienced teachers, but we have not gathered complementary data from beginning teachers directly. Hence, it would be useful for further research to use larger samples of teachers in schools spread over the four clusters.

Fourth, the scope of this study was narrowed down to the interpersonal aspect of PLCs for the cluster analysis. Future studies could be directed at providing a broader picture, which takes elements of personal and organizational variables into account (Sleegers et al., 2013).

Despite these limitations, we think that our mixed method design offers several opportunities of future research in school improvement. A main advantage of our design is that it provides a method of identifying contrasting cases in interpersonal capacity and of better understanding why there is a difference in the interpersonal capacity between schools. An important challenge in school improvement research is the identification of different stages of school capacity. It is important to realize that schools differ in their key characteristics of what makes a school great. Our study provides a method to identify different stages in the interpersonal capacity of schools. A similar method can be used to identify different stages in other key characteristics of schools. The purposeful selection of cases provides another methodological opportunity of future school improvement research. By analyzing the data from a school perspective, the key characteristics of the study, collaboration and teacher learning, are placed in the context of the whole school. The school perspective shows how several elements are connected to each other and how their coherence results in an organizational configuration. It is precisely the specific connection between several elements that results in different forms of teacher learning at different schools. By using contrasting cases, it becomes obvious what eventually makes the difference between schools. It is more difficult to understand what really makes the difference in studies that only focus on high-performing schools. It is the comparison between high and low performing schools on specific characteristics that makes it clear, what aspects are fundamental for differences in school capacity.

Finally, we believe that our use of digital logs is an interesting method of future longitudinal research. A long-term approach provides an additional perspective to school improvement research. The analysis of how teachers perceive the evolution of school characteristics over a longer period of time, e.g. a whole school year as in our study, provides useful insights into how schools deal with innovation, how they integrate this innovation into their internal operations, and how this leads to more or fewer effects in the professional development of their teachers. We hope that these methodological reflections can be an inspiration for future school improvement research.

References

- Achinstein, B. (2002). Conflict amid community: The micropolitics of teacher collaboration. *Teacher College Record*, 104(3), 421–455.
- Bakkenes, I., Vermunt, J. D., & Wubbels, T. (2010). Teacher learning in the context of educational innovation: Learning activities and learning outcomes of experienced teachers. *Learning and Instruction*, 20(6), 533–548. <https://doi.org/10.1016/j.learninstruc.2009.09.001>

- Bolam, R., McMahon, A. J., Stoll, L., Thomas, S. M., Wallace, M., Greenwood, A. M., Hawkey, K., Ingram, M., Atkinson, A., & Smith, M. C. (2005). *Creating and sustaining effective professional learning communities*. DfES, GTCe, NCSL. <https://www.education.gov.uk/publications/eOrderingDownload/RR637-2.pdf>
- Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, 33(8), 3–15. <https://doi.org/10.3102/0013189X033008003>
- Bryk, A. S., Camburn, E., & Louis, K. S. (1999). Professional community in Chicago elementary schools: Facilitating factors and organizational consequences. *Educational Administration Quarterly*, 35(5), 751–781. <https://doi.org/10.1177/0013161X99355004>
- Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, 18(8), 947–967. [https://doi.org/10.1016/S0742-051X\(02\)00053-7](https://doi.org/10.1016/S0742-051X(02)00053-7)
- Clement, M., & Vandenberghe, R. (2000). Teachers' professional development: A solitary or collegial (ad)venture? *Teaching and Teacher Education*, 16(1), 81–101. [https://doi.org/10.1016/S0742-051X\(99\)00051-7](https://doi.org/10.1016/S0742-051X(99)00051-7)
- Creswell, J. W. (2008). *Educational research. Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Pearson Education.
- Darling-Hammond, L., Chung Wei, R., Alethea, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Stanford, CA: National Staff Development Council and The School Redesign Network.
- Deakin Crick, R. (2008). Pedagogy for citizenship. In F. Oser & W. Veugelers (Eds.), *Getting involved: Global citizenship development and sources of moral values* (pp. 31–55). Rotterdam, The Netherlands: Sense.
- Desimone, L. (2009). Improving impact studies of teachers' professional development: Towards better conceptualizations and measures. *Educational Researcher*, 38(3), 181–199. <https://doi.org/10.3102/0013189X08331140>
- Doppenberg, J. J., Bakx, A. W. E. A., & den Brok, P. J. (2012). Collaborative teacher learning in different primary school settings. *Teachers and Teaching: Theory and Practice*, 18(5), 547–566. <https://doi.org/10.1080/13540602.2012.709731>
- Eraut, M. (2004). Informal learning in the workplace. *Studies in Continuing Education*, 26(2), 247–273. <https://doi.org/10.1080/158037042000225245>
- Gore, P. A. (2000). Cluster analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 297–321). San Diego, CA: Academic.
- Greene, J. C., Caracilli, V. J., & Graham, W. F. (1989). Towards a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(3), 255–274. <https://doi.org/10.3102/01623737011003255>
- Hammerness, K., Darling-Hammond, L., Bransford, J., Berliner, D. C., Cochran-Smith, M., McDonald, M., & Zeichner, K. (2005). How teachers learn and develop. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world* (pp. 358–389). San Francisco, CA: Jossey-Bass.
- Hargreaves, A. (1994). *Changing teachers, changing times: Teachers' work and culture in the postmodern age*. London, UK: Cassell.
- Hargreaves, A. (2000). Four ages of professionalism and professional learning. *Teachers and Teaching: Theory and Practice*, 6(2), 151–182. <https://doi.org/10.1080/713698714>
- Hipp, K. K., Huffman, J. B., Pankake, A. M., & Olivier, D. F. (2008). Sustaining professional learning communities: Case studies. *Journal of Educational Change*, 9(2), 173–195. <https://doi.org/10.1007/s10833-007-9060-8>
- Hoekstra, A., Brekelmans, M., Beijaard, D., & Korthagen, F. (2009). Experienced teachers' informal learning: Learning activities and changes in behavior and cognition. *Teaching and Teacher Education*, 25, 663–673. <https://doi.org/10.1016/j.tate.2008.12.007>
- Hord, S. M. (1986). A synthesis of research on organizational collaboration. *Educational Leadership*, 43(5), 22–26.

- Hord, S. M. (1997). *Professional learning communities: Communities of continuous inquiry and improvement*. Austin, TX: Southwest Educational Development Laboratory.
- Huberman, M. (1989). On teachers' careers: Once over lightly, with a broad brush. *International Journal of Educational Research*, 13(4), 347–362. [https://doi.org/10.1016/0883-0355\(89\)90033-5](https://doi.org/10.1016/0883-0355(89)90033-5)
- Johnson, B. (2003). Teacher collaboration: Good for some, not so good for others. *Educational Studies*, 29(4), 337–350. <https://doi.org/10.1080/0305569032000159651>
- Kwakman, K. (2003). Factors affecting teachers' participation in professional learning activities. *Teaching and Teacher Education*, 19(2), 149–170. [https://doi.org/10.1016/s0742-051x\(02\)00101-4](https://doi.org/10.1016/s0742-051x(02)00101-4)
- Leech, N. L., & Onwuegbuzie, A. J. (2009). A typology of mixed methods research designs. *Quality & Quantity*, 43(2), 265–275. <https://doi.org/10.1007/s11135-007-9105-3>
- Lieberman, A., & Pointer Mace, D. H. (2008). Teacher learning: The key to educational reform. *Journal of Teacher Education*, 59(3), 226–234. <https://doi.org/10.1177/0022487108317020>
- Little, J. W. (1990). The persistence of privacy - autonomy and initiative in teachers professional relations. *Teachers College Record*, 91(4), 509–536.
- Lomos, C., Hofman, R. H., & Bosker, R. J. (2011). The relationship between departments as professional communities and student achievement in secondary schools. *Teaching and Teacher Education*, 27(4), 722–731. <https://doi.org/10.1016/j.tate.2010.12.003>
- Louis, K. S., Dretzke, B., & Wahlstrom, K. (2010). How does leadership affect student achievement? Results from a national US survey. *School Effectiveness and School Improvement*, 21(3), 315–336. <https://doi.org/10.1080/09243453.2010.486586>
- Louis, K. S., & Marks, H. M. (1998). Does professional community affect the classroom? Teachers' work and student experience in restructuring schools. *American Journal of Education*, 106, 532–575.
- McLaughlin, M. W., & Talbert, J. E. (2001). *Professional communities and the work of high school teaching* (2nd ed.). Chicago, IL: University of Chicago Press.
- Meirink, J. A., Imants, J., Meijer, P. C., & Verloop, N. (2010). Teacher learning and collaboration in innovative teams. *Cambridge Journal of Education*, 40(2), 161–181. <https://doi.org/10.1080/00305764X.2010.481256>
- Meirink, J. A., Meijer, P. C., & Verloop, N. (2007). A closer look at teachers' individual learning in collaborative settings. *Teachers and Teaching: Theory and Practice*, 13(2), 145–164. <https://doi.org/10.1080/13540600601152496>
- Miles, M., & Huberman, M. (1994). *Qualitative data analysis*. London, UK: Sage.
- Newmann, F. M., Marks, H. M., Louis, K. S., Kruse, S. D., & Gamoran, A. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. San Francisco, CA: Jossey-Bass.
- OECD. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. Paris, France: OECD Publishing.
- Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage.
- Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about teacher learning? *Educational Researcher*, 29(1), 4–15. <https://doi.org/10.3102/0013189X029001004>
- Richter, D., Kunter, M., Klusmann, U., Lüdtke, O., & Baumert, J. (2011). Professional development across the teaching career: Teachers' uptake of formal and informal learning opportunities. *Teaching and Teacher Education*, 27(1), 116–126. <https://doi.org/10.1016/j.tate.2010.07.008>
- Scribner, J. S. (1999). Professional development: Untangling the influence of work context on teacher learning. *Educational Administration Quarterly*, 35(2), 238–266. <https://doi.org/10.1177/0013161X99352004>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. 86, 420–428. doi:<https://doi.org/10.1037/0033-2909.86.2.420>
- Sleegers, P., den Brok, P., Verbiest, E., Moolenaar, N. M., & Daly, A. J. (2013). Towards conceptual clarity: A multidimensional, multilevel model of professional learning communities in Dutch elementary schools. *The Elementary School Journal*, 114(1), 118–137. <https://doi.org/10.1086/671063>

- Stoll, L., Bolam, R., McMahon, A., Wallace, M., & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of Educational Change*, 7(4), 221–258. <https://doi.org/10.1007/s10833-006-0001-8>
- Svanbjörnsdóttir, B. M., Macdonald, A., & Frímansson, G. H. (2016). Teamwork in establishing a professional learning community in a new Icelandic school. *Scandinavian Journal of Educational Research*, 60(1), 90–109. <https://doi.org/10.1080/00313831.2014.996595>
- Tam, A. C. F. (2015). The role of a professional learning community in teacher change: A perspective from beliefs and practices. *Teachers and Teaching*, 21(1), 22–43. <https://doi.org/10.1080/013540602.2014.928122>
- van Veen, K., Zwart, R. C., Meirink, J. A., & Verloop, N. (2010). *Professionele ontwikkeling van leraren: Een reviewstudie naar effectieve kenmerken van professionaliseringsinterventies van leraren* [Teachers' professional development: A review study on effective characteristics of professional development initiatives for teachers]. Leiden, The Netherlands: ICLON.
- Vanblaere, B., & Devos, G. (2016). Exploring the link between experienced teachers' learning outcomes and individual and professional learning community characteristics. *School Effectiveness and School Improvement*, 27(2), 205–227. <https://doi.org/10.1080/09243453.2015.1064455>
- Vandenbergh, R., & Kelchtermans, G. (2002). Leraren die leren om professioneel te blijven leren: Kanttekeningen over context [Teachers learning to keep learning professionally: Reflections on context]. *Pedagogische Studiën*, 79, 339–351.
- Vangrieken, K., Dochy, F., Raes, E., & Kyndt, E. (2015). Teacher collaboration: A systematic review. *Educational Research Review*, 15(1), 17–40. <https://doi.org/10.1016/j.edurev.2015.04.002>
- Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education*, 24(1), 80–91. <https://doi.org/10.1016/j.tate.2007.01.004>
- Visscher, A. J., & Witziers, B. (2004). Subject departments as professional communities? *British Educational Research Journal*, 30(6), 786–801. <https://doi.org/10.1080/0141192042000279503>
- Wahlstrom, K., & Louis, K. S. (2008). How teachers experience principal leadership: The roles of professional community, trust, efficacy, and shared responsibility. *Educational Administration Quarterly*, 44(4), 458–496. <https://doi.org/10.1177/0013161X08321502>
- Wang, T. (2015). Contrived collegiality versus genuine collegiality: Demystifying professional learning communities in Chinese schools. *Compare: A Journal of Comparative and International Education*, 45(6), 908–930. <https://doi.org/10.1080/03057925.2014.952953>
- Zwart, R. C., Wubbels, T., Bergen, T., & Bolhuis, S. (2009). Which characteristics of a reciprocal peer coaching context affect teacher learning as perceived by teachers and their students? *Journal of Teacher Education*, 60(3), 243–257. <https://doi.org/10.1177/0022487109336968>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 11

Recurrence Quantification Analysis as a Methodological Innovation for School Improvement Research



Arnoud Oude Groote Beverborg, Maarten Wijnants, Peter J. C. Slegers,
and Tobias Feldhoff

11.1 Introduction

In educational research and practice, teacher learning in schools is recognized as an important resource in support of school improvement and educational change. In their efforts to understand the mechanisms underlying school improvement, researchers have started to examine the role of teacher learning as a key component to building school-wide capacity to change. In practice, professional learning communities are being increasingly developed to stimulate the sharing of knowledge, information and expertise among teachers, with the goal to improve instruction and student learning. More specifically, by engaging in professional learning activities, teachers can make knowledge and information explicit, discover the proper scripts for future actions aimed at adaptation to changes such as ongoing reorganizations of work processes and accountability reforms, and to formulate and monitor goals for further development of for instance instructional methods and technological innovations (Korthagen, 2010; Oude Groote Beverborg, Slegers, Endedijk, & van Veen, 2015a).

To understand more about how engagement in professional learning activities enables teachers to learn, scholars have called for more situated and longitudinal research (Feldhoff, Radisch, & Bischof, 2016; Feldhoff, Radisch, & Klieme, 2014; Korthagen, 2010). The few longitudinal studies conducted so far used analytic

A. Oude Groote Beverborg (✉) · M. Wijnants
Radboud University Nijmegen, Nijmegen, The Netherlands
e-mail: a.oudegrootebeverborg@fm.ru.nl

P. J. C. Slegers
BMC, Amersfoort, The Netherlands

T. Feldhoff
Johannes Gutenberg University, Mainz, Germany

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_11

219

techniques (Structural Equation Modelling; SEM) that derive their power from large samples of participants and included a limited number of measurement occasions with relatively long intervals (e.g. yearly intervals) to assess the (reciprocal) relationships between variables under study. The findings suggest, among other things, that reflection is positively related to self-efficacy and changes in instructional practices (Oude Groote Beverborg, et al., 2015a; Slegers, Thoonen, Oort, & Peetsma, 2014). Higher levels of engagement in professional learning activities, thus, seem beneficial to improve education. In addition, these studies pointed towards the importance of conditions at the school-level, such as transformational leadership and working in teams, to foster teacher learning. This suggests that a purposeful and empowering environment can help to structure uncertainty and ambiguity, and to enable teachers to come to a common understanding about changing their practice, and learn from one another (see also Coburn, 2004; Oude Groote Beverborg, 2015; Staples & Webster, 2008). As such, these longitudinal studies have their merit in validating and extending previous findings from cross-sectional studies on the structural relations between organizational conditions and improving education over time (see also Hallinger & Heck, 2011; Heck & Hallinger, 2009; Heck & Hallinger, 2010).

However, findings on structures at the school-level do not inform about how teachers use these organizational conditions in everyday regulation practices and how such use may fluctuate over time (Maag Merki, Grob, Rechsteiner, Rickenbacher, & Wullschleger, 2021, see chapter 12; see also Hamaker, 2012; Molenaar & Campbell, 2009). It remains for instance unclear how higher levels of engagement in professional learning activities translate to individual teachers' routines of for instance reflection or knowledge sharing on a daily basis (see also Little & Horn, 2007). Are these higher levels based on for instance reflecting very regularly (every day a little) or in bursts (whenever there is a necessity or opportunity)? By extension, it remains unclear whether the regularity with which moments of teacher learning are organized also contributes to sustaining school improvement (think with regard to regularity for instance of the rhythm of reflection cycle phases for self-improvement, the periodicity of meetings of learning community members to develop instruction and curriculum, and even the intervals of appraisal interviews and classroom observations that can be used for quality development monitoring and accountability purposes) (e.g. Desimone, 2009; Korthagen, 2001; van der Lans, 2018; van der Lans, van de Grift, & van Veen, 2018).

In contrast to large survey studies, case studies have generated situated descriptions of what occurs during efforts to improve schools in specific contexts (see for instance Coburn, 2001, 2005, 2006). However, case studies do not have the aim to generalize their findings, and the validity and utility of those findings is limited. As such, the available research provides no systematic evidence of how (for what and when) teacher learning takes shape in its social context. Consequently, understanding more about the dynamics of everyday teacher learning and its link with school improvement and educational change requires studies that are situated, longitudinal,

and aimed at finding systematic relations, and in addition, a corresponding situated and dynamic perspective (Barab et al., 1999; Clarke & Hollingsworth, 2002; Greeno, 1998; Heft, 2001; Horn, 2005; Lave & Wenger, 1991; Reed, 1996).

From a situated and dynamic perspective, school improvement is seen as an ongoing, embedded, complex, and dynamic process of adapting to continuously changing challenges that arise out of schools' unique circumstances. School improvement emerges from the many interactions between actors within and outside schools, making the school improvement journey highly context-sensitive, and the occurrence of meaningful developments (or milestones) unpredictable (van Geert & Steenbeek, 2014; see also Ng, 2021, chapter 7). Similarly, teacher learning is seen as a cyclical process in which available environmental information, professional learning activities, and productive practices are interconnected and co-develop (Barab et al., 1999; Clarke & Hollingsworth, 2002), that is, teachers attend to, interpret, adapt, and transform information from their environment and make use of their (social) environment to learn what is needed (Barab & Roth, 2006; Gibson, 1979/1986; Greeno, 1998; Little, 1990; Maitlis, 2005).

Investigating ongoing micro-level change processes, such as the routine with which individual teachers make environmental information and changes in meaning, knowledge, or accommodation of teaching practices, explicit, requires analytic techniques that assess intra-individual variability over time, such as State Space Grid analysis (Granic & Dishion, 2003; Lewis, Lamey, & Douglas, 1999; Mainhard, Pennings, Wubbels, & Brekelmans, 2012) or Recurrent Quantification Analysis (RQA). In contrast to commonly used statistical modelling techniques, such as SEM, these techniques are based on dense time-series, whose temporal structures are kept intact. They provide measures about the stability or flexibility of a developmental process. RQA has been applied to analyse coordination in conversations, reading fluency, emergence of insights and behavioural changes (Dale & Spivey, 2005; Lichtwarck-Aschoff, Hasselman, Cox, Pepler, & Granic, 2012; O'Brien, Wallot, Haussmann, & Kloos, 2014; Richardson, Dale, & Kirkham, 2007; Stephen, Dixon, & Isenhower, 2009; Wijnants, Hasselman, Cox, Bosman, & Van Orden, 2012; see also Wijnants, Bosman, Hasselman, Cox, & Van Orden, 2009).

This study aims to examine the overall level and the routine of learning through reflection in the workplace. More specifically, this study focusses on the relation between the temporal pattern of becoming aware of information in the (social) environment and experiencing new insights by making both explicit through reflection. It does so by collecting dense intra-individual (teacher) longitudinal measurements (logs), and by illustrating how RQA can be applied to these time-series. We will explore the application of RQA as a promising analytic technique for understanding the co-evolution of teacher learning and school-wide capacity for sustained improvement.

11.2 Theoretical and Methodological Framework

In this section, we will first describe teachers as active interpreters of their specific circumstances and as reflective practitioners (e.g. Clarke & Hollingsworth, 2002). Next, we will discuss and describe logs as measurement instruments that can capture this situated process over time. Thereafter, we will extensively discuss RQA and we will present examples of studies to provide some research context as to how it can be applied. We will end this section by showing how this conceptualization, measurement instrument, and analysis strategy come together in the present study.

11.2.1 *Information and Reflection in a Situated and Ongoing Learning Process*

Within the situated perspective, teacher learning is considered an acculturation process (Greeno, 1998; Lave & Wenger, 1991). Teachers are considered active, intentional perceivers, constructing a meaningful practice by integrating new experiences with old experiences (Coburn, 2004; Slegers & Spillane, 2009; Spillane & Miele, 2007). These experiences are provided by the community while the person is engaged in it (Lave & Wenger, 1991; Little, 2003; Wenger, 1998). Central to this perspective is that knowledge is distributed over a situation (Greeno, 1998; Hutchins, 1995; Putnam & Borko, 2000), that a person makes sense of it through action (Little, 2003; Spillane, Reiser, & Reimer, 2002; Weick, 2011), and that sensemaking is embedded in a person's history (Coburn, 2001; Coburn, 2004; Slegers, Wassink, van Veen, & Imants, 2009), as well as in a social and cultural context (Slegers & Spillane, 2009). While acting, a person selects the information that affords continued action and that fits the understanding of the purpose in the situation (Coburn, 2001; Slegers et al., 2009; Spillane et al., 2002). Learning can thereby also be characterized as a process of continuously attuning (Barab et al., 1999; Clarke & Hollingsworth, 2002; Granic & Dishion, 2003; Guastello, 2002). As such, teachers can regulate what information in the (social) environment they attend to, so that, over a longer period of time, experiences of interactions with the (social) environment consolidate into new, or differentiations of, meanings, knowledge, and skills (Korthagen, 2010; Kunnen & Bosma, 2000; Lichtwarck-Aschoff, Kunnen, & van Geert, 2009; Steenbeek & van Geert, 2007; van Geert & Steenbeek, 2005). In addition, of course, teachers can develop and adapt by regulating their activities through reflection (Argyris & Schön, 1974; Korthagen & Vasalos, 2005; Schön, 1983).

Teacher engagement in reflection, then, can be seen as an introspective activity that refers to a person recreating an experience of acting in a given situation. In making this experience explicit later, a person supplements the memory of the experience with new ideas that can either be self-generated or based on information gained from others (Oude Groote Beverborg, Slegers, & van Veen, 2015b). This creates an altered and thus new experience, which can then serve as the basis for

future action. In this way, reflection directs what information in the environment is to be attended to, thought about, and reacted to, and for what purpose (Clarke & Hollingsworth, 2002; see also Weick, 2006). Making information explicit in this way helps to put the knowledge that is distributed within teachers' environments to focussed use and regulates development and adaptation by setting priorities for attention and actions. As such, making previously encountered information explicit shapes future experiences, what can be consequently reflection upon, and what will be made explicit thereafter. This interplay between environmental information and reflection stresses that the directions teachers' and their school's developments can take are based in a teacher's specific circumstances.

Moreover, through repeated investigation of one's own actions and encountered information, a teacher might, after a while, suddenly discover a new way of acting or looking at the world that is more functional in a given situation than the old one was (Clarke & Hollingsworth, 2002). Such learning experiences of change in meaning, knowledge, or skills, which were generated by one person, can also be reflected upon, made explicit, and shared as possibly of value for other individuals and the team (Nonaka, 1994; van Woerkom, 2004). That also helps to find solutions to ongoing changes and challenges at work, and to formulate and monitor goals for further development (of for instance shared meaning) and improvement (of for instance a school's capacity for change) (Oude Groote Beverborg, et al., 2015a).

However, due to the circumstantial and temporal dependency of available information, meaning, knowledge, and skills, intensities of engagement in reflection on one's working environment can fluctuate over time within persons and can differ between persons before new insights emerge (Endedijk, Brekelmans, Verloop, Slegers, & Vermunt, 2014; Stephen & Dixon, 2009; see also Orton & Weick, 1990). The corresponding trajectories of individual teachers' engagements in making information explicit may therefore look quite irregular and not alike. Additionally, learning experiences can also emerge with different intervals. Repeated engagement in reflection on one's working environment therefore changes, continuously slightly (sensitivity to specific information) and occasionally more profoundly (experience of having learned something), the way the world is perceived, understood, and enacted (see also Coburn, 2004; Voestermans & Verheggen, 2007, 2013).

Nevertheless, it remains unclear with how much routine teachers engage in reflection in their everyday practices. Insights into the intra-individual variability in intensity of everyday reflection may provide valuable knowledge to schools as well as to the inspectorates of education about the ways, in which they can organize and support teacher learning in the workplace. In order to tap into these dynamics of reflection and their consequences, measurement instruments therefore need to be designed that allow for specific person-environment interactions and that can be administered densely (see also Bolger & Laurenceau, 2013). Moreover, the chosen analysis needs to provide measures that can represent temporal variability. In the next two sections, we will address the use of logs as a measurement instrument that can be administered densely and the use of RQA as an analytic technique that yields dynamics measures.

11.2.2 Logs

In order to tap into the dynamics of individual teachers' reflection processes, it is necessary to look at them while and where they are happening – rather than by means of for instance interviews that are prone to hindsight bias or with standardized questionnaires that are insensitive to specific circumstances – to focus on the continuous interaction between the acting professional and the environment through time, and then reconstruct the learning process as a series of interactions over time (see for an example Endedijk, Hoekman, & Slegers, 2014; Lunenberg, Korthagen, & Zwart, 2011; Lunenberg, Zwart, & Korthagen, 2010; Zwart, Wubbels, Bergen, & Bolhuis, 2007; Zwart, Wubbels, Bolhuis, & Bergen, 2008). This would give an account of professional development including prospective learning, and not only an account of retrospective learning.

In this study, we will therefore measure teachers' reflection processes with logs (for other uses of logs in dynamic analyses, see: Guastello, Johnson, & Rieke, 1999; Lichtwarck-Aschoff et al., 2009; Maitlis, 2005; for other uses of logs in school improvement research, see Maag Merki et al., 2021, chapter 12; Spillane & Zuberi, 2021, chapter 9). Not everything that happens can be reported in a log. What is reported, is what is most salient in a teacher's experience. Using open questions, this can be charted in a personalized and situated manner.

The use of logs presupposes that teachers have a sensitivity to information in their environment, that they monitor their development, and that they have an affinity for making information and knowledge explicit by using logs. Every time teachers fill in a log entry, they use an opportunity to make information, experiences, or knowledge explicit (as, in a sense, surveys with targeted items and interviews with targeted questions do as well). Participating in this study might therefore make teachers more aware of what is going on in their environment, of their purpose, and in what areas they develop (Geursen, de Heer, Korthagen, Lunenberg, & Zwart, 2010). By administering logs densely, the logs themselves can also become a familiar part of the working environment that teachers can choose to engage with. Nevertheless, teachers flow with the issues of the day, and may find it hard to disengage from the immediacy of their work to make time to reflect by using logs. Logs thereby not only measure the learning process. They do so by setting a model of the reflection process in terms of content and pace that may fit better or worse to different teachers within a certain period of time. Moreover, the interval with which logs are administered ought to be in accord with the expected rate of change of the frequency with which teachers are likely to reflect upon their environment and learning experiences.

For the assessment of reflection routines, it is important that logs can generate a dense time-series. From these time-series, the dynamics of engagement in reflection can be reconstructed with an RQA.

11.2.3 *Recurrence Quantification Analysis*

RQA is a nonlinear technique to quantify recurring patterns and parameters pertaining to the stability of the underlying dynamics from a time-series (with an intact temporal structure). An important advantage of RQA, unlike other time-series analysis methods, is that this technique does not impose constraints on data-set size (N). RQA does not make assumptions regarding statistical distributions or stationarity of data either. Nevertheless, for RQA to provide interpretable results, it has been suggested that the minimum requirements for the time-series are that they are long enough to contain at least two repetitions of the whole repeating dynamic pattern and that at least three measurement occasions fall within each repetition of the repeating dynamic pattern (Brick, Gray, & Staples, 2018). Needless to say that more robust and precise estimation will be permitted by measuring longer and denser, which may thus be required for noisier data. The technique reveals subtle time-evolutionary behaviour of complex systems by quantifying system characteristics that would otherwise have remained hidden (i.e., when only taking frequencies into account). To get an idea of what is meant by dynamics, consider Fig. 11.1. It shows five examples of hypothetical, idealized change trajectories (i.e. stability, growth, randomness, and two times regular fluctuation) of engagement in reflection of different persons. Trajectories (**a**, **b**, **c**, and **d**) all have different temporal patterns (rhythms). Their overall level of reflection does not distinguish them: Each trajectory has a mean of 1. In comparison, trajectories (**d** and **e**) differ in their means, but have the same rhythm. The differences between the change trajectories become apparent, because they have (relatively) many time-points.

A distinction can be made between the application of RQA to categorical (nominal) data¹ and to continuous (scale) data. Categorical RQA is a simplified form of continuous RQA². This chapter will focus on categorical RQA. Moreover, RQA can be applied to single time-series (auto-RQA) or to two different time-series (cross-RQA). Fundamentally, auto-RQA is applied to answer questions concerning

¹RQA allows a direct access to dynamic systems (characterized by a large number of participating, often interacting variables) by reconstructing, from a single measured variable in the interactive system, a behaviour space (or phase-space) that represents the dynamics of the entire system. This reconstruction is achieved by the method of delay-embedding that is based on Takens' theorem (Broer & Takens, 2009; Takens, 1981). The phase space reconstructed from the time series of this single variable informs about the behaviour of the entire system because the influence of any interdependent, dynamical variable is contained in the measured signal. The reconstruction itself involves creating time-delayed copies of the time-series of a variable that become the surrogate dimensions of a multi-dimensional phase-space. Consequently, the original variable becomes a dimension of the system in question and each time-delayed copy becomes another dimension of the system. Because of that, it is not needed to know all elements of the system, or measure them, to reconstruct the behaviour of a dynamic system, provided that a (sufficiently dense) time-series of one element of the system is available. For tutorials on continuous RQA, see: Marwan et al. (2007) and Riley and Van Orden (2005). For applications of continuous RQA in the social sciences, see: Richardson, Schmidt, and Kay (2007) and Shockley, Santana, and Fowler (2003).

²Delay-embedding is not applied – the system is considered to have 1 dimension.

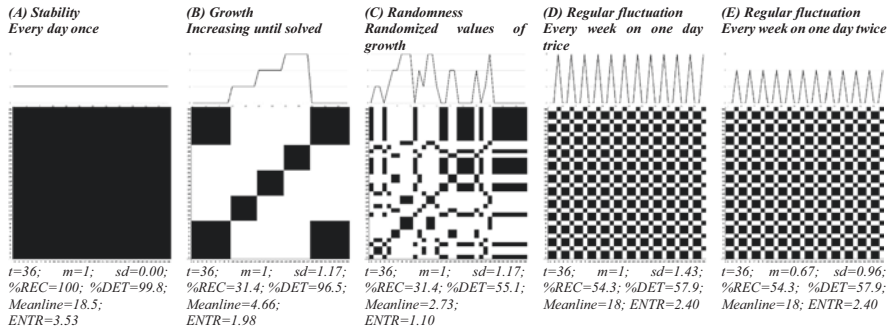


Fig. 11.1 Five examples of change trajectories, shown as time-series graphs and recurrence plots, of engagement in reflection with different dynamics

Note: Change trajectories (a, b, c, d and e) represent hypothetical, idealized change trajectories (i.e. stability, growth, randomness, and two times regular fluctuation, respectively) of engagement in reflection of different persons. Trajectories (a, b, c and d) all have a mean of 1 but differ in the values of their dynamics (rhythm) measures. In comparison, trajectories (d and e) differ in their means, but have the same values of their dynamics measures. Each trajectory is represented by two graphs: one time-series and one recurrence plot (top and bottom graphs, respectively). The time-series have 36 time points (i.e. days) (x-axis of each graph) and engagement in reflection can have one of the following values at each time point: 0, 1, 2, or 3 (i.e. the number of reflection moments, or the amount of reflection intensity, per day) (y-axis of each graph). In the recurrence plots, both the x-axis and the y-axis represent the 36 time points, and thus the plots have $36 \times 36 = 1296$ cells. These cells can either be filled or empty (filling is in this case marked by a black square). Filled cells are called recurrence points. Recurrence points represent that the process had a value at a certain time point and that that value also occurred at another time point (i.e. the recurrence of one of the reflection intensity values). In these examples, the time-series are plotted against themselves in the recurrence plots (i.e. auto-recurrence), and thus the plots are symmetrical around the Line of Incidence (the center diagonal line, i.e. the time-series as it was measured). Auto-recurrence plots are generated for each single time-series separately. The Line of Incidence is excluded in the calculation of the dynamics measures. t = length of the time-series; m = mean of the values in the time-series; sd = standard deviation around the mean; $\%REC$ = Recurrence Rate (i.e. the percentage of recurrence points in the recurrence plot); $\%DET$ = Determinism (i.e. the percentage of recurrence points that form diagonal lines out of the total of recurrence points); $Meanline$ = the mean length of all diagonal lines of recurrence points; $ENTR$ = Shannon Entropy (i.e. a measure of complexity; it is calculated as the sum of the probability of observing a diagonal Line Length times the log base 2 of that probability). See also the Recurrence Quantification Analysis-section and Fig. 11.3

within-actor variability, whereas cross-RQA is applied to answer questions concerning variability in coordination between actors over time.

RQA combines the visualization of temporal dynamics in recurrence plots with the objective quantification of (non-linear) system properties. In auto-RQA, one time-series is placed on both the x-axis and the y-axis to generate the recurrence plot. In cross-RQA, one time-series is placed on the x-axis and another time-series is placed on the y-axis to generate the recurrence plot. In essence, a recurrence plot is a graphical representation of a binomial matrix that shows after what delays

values in time-series recur (recurrence points³). The recurrence plot is then quantified and used to calculate complexity measures.

Consider Fig. 11.1 again. In the figure, engagement in reflection has one of the following values at each time point: 0, 1, 2, or 3 (i.e. the number of reflection moments, or the amount of reflection intensity, per day). The temporal order of these values is given in the time-series graphs. The recurrence plots on the other hand are composed of auto-recurrence points; that is, they show that any of these values occurred at a certain moment and that that also happened sometime else within the same time-series (earlier, at the same time, or later). In these examples, the time-series are plotted against themselves in the recurrence plots (i.e. auto-recurrence), and thus the plots are symmetrical around the Line of Incidence (the centre diagonal line, i.e. the actual time-series – in cross-RQA, this line is sometimes called the Line of Synchrony). Auto-recurrence plots are generated separately for each single time-series. The time-series graph of the stable process in (a) shows that at each time point the process had a value of 1. Therefore, the corresponding recurrence plot is fully filled. In comparison, the growth (and decline) process in (b) shows a steady increase from 0 to 3 followed by a sharp decrease to 0 again. Consequently, the recurrence plot shows neatly clustered recurrence points. The random process in (c) has the same time-series values as the time-series in (b), but in (c), the temporal structure of these values was changed by placing them in a random order. Consequently, the recurrence plot of the process in (c) is less characterized by diagonal lines (consecutive recurrences form diagonal lines). Therefore, the process in (c) has the same values as in (b) for the mean and the Recurrence Rate, but the other dynamics measures differ. The regularly fluctuating processes in (d and e) both have only two values (0 and 3, or 0 and 2, respectively), and in both trajectories, these values recur after the same period. Therefore, they have identical recurrence plots and thus identical dynamics measures.

When the same behaviour is repeated periodically or when different behaviours succeed each other periodically, diagonal lines are formed in the recurrence plot. Measures based on the temporal order of these recurrence-sequences in the recurrence plot inform about the dynamics of the system. The Line of Incidence is excluded in the calculation of the dynamics measures. We will introduce the measures Recurrence Rate, Determinism, Meanline, and Entropy (other measures are Maxline, Laminarity, and Trapping Time) (Marwan, Romano, Thiel, & Kurths, 2007; see also Cox, van der Steen, Guevara, de Jonge-Hoekstra, & van Dijk, 2016) and elaborate on three studies as examples of how to apply them.

Recurrence Rate is computed as the ratio of the number of recurrent points (the black regions in the recurrence plot) over the total number of possible recurrence points in the recurrence plot (i.e. the length of the time-series squared). The Recurrence Rate thus indicates how often behaviours in a time-series re-occur (or also occur in the case of cross-RQA). The Recurrence Rate is not based on the

³Note that for categorical RQA, values need to be clearly demarcated categories to form recurrence points.

temporal order of the values in the time-series, and is thus a raw measure of variability of behaviour (or of coordination in the behaviours of two actors in the case of cross-RQA) over time.

Determinism is defined as the ratio of the number of recurrence points forming a diagonal pattern (i.e. a sequence of recurring behaviours) over the total number of recurrence points in the recurrence plot. Determinism thus informs about behaviours that continue to recur over time relative to isolated recurrences, indicating the persistence of those behaviours.

An example of a study using Recurrence Rate and Determinism was conducted by Dale and Spivey (2005). They applied categorical cross-RQA to assess lexical and syntactic coordination in conversations of dyads of children and caregivers at many measurement occasions ($N_{\text{dyads}} = 3$; $N_{\text{participants}} = 6$; $N_{\text{conversations}}$ were 181, 269, and 415). They used the Recurrence Rates of words and of grammar as an indication of coordination between child and caregiver. Types of words are more numerous in conversations than syntactic classes, and types of words therefore gives lower Recurrence Rate values. Additionally, they used the Determinism of words and of grammar, but now based on the set of words that lay within about 50 words from each other in the conversations (i.e. within the band of about 50 words around the Line of Synchrony). This provides an indication of dynamic structures of coordination that are closer together in time and it forms a basis for the interpretation of the Recurrence Rate. Then, they computed both measures again, but now based on the child's time-series at the same measurement occasion and the caregiver's time-series at a measurement occasion one step ahead in development. They compared the 2×2 Recurrence measures and the 2×2 Determinism measures of each dyad using t-tests to assess the influence of the given conversation. Finally, they assessed the development of the Recurrence Rate and Determinism over time using regression analyses. For all comparisons of RQA measures, results indicated that coordination between child and caregiver was stronger within the same entire conversation than over conversations, and that coordination was stronger with greater temporal proximity within a conversation. Moreover, the results indicated that coordination diminished over development.

Meanline is an index of the average duration of deterministic patterns, and thus indicates how long on average the person (or dyad in the case of cross-RQA) remains in similar behavioural states over time. Meanline provides information about the stability of behaviour.

An example of a study using Meanline was conducted by O'Brien et al. (2014). They applied continuous auto-RQA to assess stability of reading fluency of children in different grades and that of adults ($N_{\text{cohorts}} = 4$; $N_{\text{participants}} = 71$; $N_{\text{texts}} = 1$). All participants read the same text. Additionally, each participant of each cohort was randomly assigned to either a silent reading or a reading out loud condition. The researchers used Meanline as a measure of the length of recurring stretches of word-reading-times (other measures relating to other aspects of reading were also used). ANOVAs were used to compare cohorts and conditions. Moreover, they applied continuous cross-RQA to each possible combination of two time-series of the participants within each cohort and within either condition. This analysis gave

shared-Meanline values. With this measure, an assessment could be made of whether the reading dynamics of each group were more structured by the text (higher shared-Meanline) or more idiosyncratically (lower shared-Meanline), that is, whether more fluent readers are less constrained by the processing of each (subsequent) word and instead follow their own meanderings through the story to monitor their own understanding of the text. Because of concerns that using the pairwise cross-RQA metric may violate the assumption of independence of observations, the shared-Meanline values were submitted to a bootstrap procedure that drew 1000 subsamples per group, after which confidence intervals were constructed for each group. Using 99% confidence intervals, those groups, whose confidence intervals did not overlap, differed significantly from the other groups. The results indicated that adults had more stability in reading in both reading modes as compared to the other cohorts, and that, when reading out loud, the reading dynamics of both sixth graders and adults are structured more idiosyncratically than those of second and fourth graders and also than those of all cohorts during silent reading.

Entropy is computed as the Shannon Entropy of the distribution of the different lengths of the deterministic segments⁴. Entropy indicates the level of complexity of the sequences of behaviours. The Entropy measure, in RQA, thus indicates how much “disorder” there is in the duration of recurrent sequences.

In the form of peak-Entropy, Entropy can for instance be used as a measure of reorganization⁵. Lichtwarck-Aschoff et al. (2012) conducted a study on the course and effect of clinical treatment for externalizing behaviour problems of children (age-range = 7–12 years). A pattern of reorganization over the course of treatment would be an indication of improvement. Both parents and children received treatment once a week for 12 weeks. Bi-weekly 4 or 6-min observations of problem solving discussions between parent and child formed the raw data ($N_{\text{dyads}} = 41$; $N_{\text{participants}} = 82$; $N_{\text{conversations}} = 6$). The data of each participant were initially coded in real-time along nine mutually exclusive affect codes for each participant. The thus acquired time-series were collapsed into one time-series per dyad, resampled to have 72 data points, and recoded along four categories (plus a rest category) that reflected the affective state of the dyad (unordered categorical data). The researchers applied categorical auto-RQA to these dyadic time-series to calculate the Entropy of each conversation of each dyad. 15,000 bootstrap replications of the sample’s Entropy values were used to estimate 95% confidence intervals. The

⁴Shannon Entropy is calculated as the sum of the probability of observing a diagonal Line Length times the log base 2 of that probability. This measure depends therefore on the number of different lengths of diagonal lines (or bins) in a particular recurrence plot. Fewer bins and more equally distributed frequencies of diagonal Line Lengths over the bins will give lower Entropy values: less information is needed to describe the behaviour of a system.

⁵For instance, learning new knowledge or skills is a reorganization of the (learner’s) system in such a way that it becomes (locally) more adapted to its environment. Having learned something new can therefore be characterized by a drop in Entropy, which then stabilizes at this lower level. The reorganization of one’s knowledge or skills, on the other hand, is a period, in which old knowledge structures or routines are broken down (after which they are reassembled), and can thus be characterized by a short peak in Entropy (see also Stephen et al., 2009 and Stephen & Dixon, 2009).

consecutive Entropy values formed the data for subsequent Latent Class Growth Analysis. This analysis was used to identify groups based on the form of the Entropy-trajectories, that is, to distinguish between conversations that could be characterized by a higher Entropy-level followed by a drop in Entropy (i.e. peak-Entropy) and conversations that did not show this pattern of reorganization. Moreover, improvement of children's externalizing behaviour problems was independently assessed through pre- and post-treatment clinicians' ratings. Based on criteria for clinically significant improvement, these ratings were also used to divide the sample into classes: improvers and non-improvers. Consequently, the two estimates of class membership were compared. The results showed that dyads in the peak-Entropy-class belonged more frequently to the improvers-class. To assess whether this finding could be simply attributed to either a decline in frequency of negative dyadic affective states or an increase in positive dyadic affective states, the researchers additionally calculated the Recurrence Rates of each coding category of each conversation (again, 95% confidence intervals were based on 15,000 bootstrap replications). The results from a non-parametric test (Kolmogorov Smirnov test) applied to these not normally distributed data showed no differences between classes in the level of recurrence of any of the affective state categories. This indicates that it might be necessary for people to have a period of unpredictability and flux, in which they try out and explore new behaviours, to develop.

11.2.4 Present Study

To reiterate, in this study, we are interested in teacher learning through reflection in the workplace. Building on a situated and dynamic perspective, learning experiences can be seen as emerging from acting upon information in the (social) environment after a period of time. Through reflection on their working environment, teachers make information explicit. Through reflection on learning experiences, teachers make new insights (developed or adapted meanings, knowledge, and skills) explicit. By making these things explicit, teachers can share them with colleagues, put them to focussed use, and set priorities concerning what to attend to and how to act in which situation. Moreover, attending to information can occur more frequently than having new insights, and therefore reflection on the working environment can occur more frequently than reflection on learning experiences. As an example of how to investigate teacher learning through reflection as an everyday and ongoing process, we designed a study to explore the routine with which teachers engage in making information explicit, and how that, in comparison to the overall levels thereof, relates to making new insights explicit. The routine of reflecting pertains to the temporal stability of that activity, and thus its dynamics should be assessed. This requires the collection of dense time-series from individual teachers.

Our measurement instruments, measurement intervals, and analytic measures were chosen in correspondence with this conceptualization. In accord with the different expected rates of change, we chose to use daily logs to measure reflection on

the environment and monthly logs to measure reflection on learning experiences. We will explore whether these measurement instruments and measurement intervals are useful for the assessment of the dynamics of learning through reflection (see also Kugler, Shaw, Vincente, & Kinsella-Shaw, 1990).

We used the responses to the daily logs to generate time-series for each participant. Each point in these time-series represents the intensity of reflection on the environment, i.e. the number of reflection moments during a day. The analysis measures for the routine of reflection on the environment were calculated by applying a categorical auto-RQA to each time-series. Recurrence Rate was used as a raw measure of routine and informs about the overall regularity of the reflection process. Determinism was used as a measure of the persistence thereof. The analysis measures for the overall level of reflection on the environment and learning experiences were calculated by simply summing up all responses to the daily and monthly logs, respectively. To investigate the extent to which the overall level and the routine of the intensity of making information explicit co-occurs with the overall intensity of making insights explicit, we generated and inspected scatterplots.

11.3 Method

We used a longitudinal, mixed-method design with convenience sampling to assess the relation between the level and routine of teachers' engagement in reflection on their environments to make information explicit and the level of reflection on learning experiences to make insights explicit. To do so, we asked teachers to fill in daily and monthly logs, including open questions about the salient information they attended to and the learning experiences they had, respectively, for a period of 5 months. Analyses were applied to the time-series of frequencies of filled in log entries.

11.3.1 Sample

This study was conducted in one VET college in the Netherlands in 2011 (see also Oude Groote Beverborg et al., 2015a). Team leaders were asked whether team members were willing to participate in this study, and participation was voluntary. A total of 20 teachers participated. The data from 1 teacher were excluded from the analysis, because the teacher had moved to a different employer (a college offering professional education), and the data from 2 other teachers were excluded, because they started 2 months late. Thus, the effective sample size was 17. The participants were employed in departments that taught law, business administration, ICT, laboratory technology, and engineering to students and that coached other teachers. Thirteen participants were female, and 4 were male. Working days per week ranged from 2 to 5. In order to generate enough data for a substantive time-series, but as a

trade-off between practicality and rigor, the study ran for 5 months: from February until June. During this period, all participants had a 2 weeks' holiday. One participant (P12) stopped participating after 2 months, and another participant (P10) after 3 months.

11.3.2 Measurement

The study consisted of two logs: a daily and a monthly log. The daily log (diary) asked teachers to make salient information explicit, and thus measured their engagement in reflection on the environment. The monthly log asked teachers to make their insights explicit, and thus measured their reflections on learning experiences. The logs were designed as short, structured interviews with a few open questions. Thereby, participants could report the information that was most relevant to them individually at each measurement occasion. More specifically, the diaries asked about the most salient information that day and the context, in which the information was attended to. The diary questions were focussed on information from colleagues (de Groot, Endedijk, Jaarsma, Simons, & van Breukelen, 2014). The main diary question was: "What did your colleague say or do that was most salient today?" It was made explicit that this could be something someone said, someone did, something that was read, and so on. Other open questions related to the task the participants worked on for which the reported information was relevant, and to how they responded to the information (see Appendix A for the complete specification of one diary entry translated into English). The diaries were designed in such a way that teachers could report their own experiences. The diaries were therefore sensitive to local and personal circumstances and measured with such a density that fluctuations could be expected to be measurable. The monthly logs were designed similarly and asked to report the learning experiences participants had had sometime in the last month as accurately as possible (Endedijk, 2010). The most important question was: "What have you learnt in the last month?" Additionally, questions about the context the learning experience came from, or in which context it had to be understood, were asked, such as about the task and the goal they related to, what means helped to learn it, the manner in which it was learnt, and in what manner participants realized they had learnt something. Lastly, the monthly log also asked questions about what teachers were satisfied with in their learning process and what could be improved in the future, what goals they would pursue in the future, and what they would attend to in the future (see Appendix B for the full specification of one monthly log entry translated into English).

Diaries were administered on each person's working days and monthly logs on the first working day of the new month. In order to constrain the burden of repeatedly filling in logs, a maximum of three diary entries (making information explicit) and three monthly log entries (making insights explicit) could be filled in per measurement occasion. Also, participants were instructed to spend no more than 5 min on each diary entry (thus a maximum of 15 min per day), and no more than 10 min

on each monthly log entry (thus a maximum of 30 min per month). Teachers were asked to fill in at least one log entry per measurement occasion, but this was not mandatory. Logs were administered online. For each participant's measurement occasion's log, an invitation was sent by email. On some measurement occasions, some invitations failed to be sent. See Fig. 11.2 and Table 11.1 for frequencies of reporting and descriptives. The analyses were applied to the time-series of frequencies of filled in log entries.

In order to uphold motivation, the first author offered individual coaching sessions to the participants. These sessions took place once every month, lasted about 45 min, and were conducted over the telephone. In general, during a session, the information a participant had reported in the log was summarized, and the participant was asked to respond to that. Towards the end of the conversation, the first author categorized some of the information in the diaries and labelled this summary, after which there was opportunity for the participant to reflect upon the labelling of the information. Each conversation ended with the first author asking feedback on the instrument and the conversation. These calls were not intended as part of the measurement of the study and have therefore not been recorded.

11.3.3 *Analysis Strategy*

The aim of the analyses was to assess in which way the overall level and the routine of the intensity of making information explicit relates to the overall intensity of making insights explicit. We calculated one measure for making insights explicit: each participant's mean of moments of reflection on learning experiences over the measurement period per month participated (overall insight intensity). This measure is based on the monthly log data. The mean per month was calculated to correct for differences between participants in the duration that they participated.

Crucially, this measure was also used to assess whether participants had affinity for the measurement instruments, that is, whether teachers disengaged from the immediacy of their work to make time to 'interact' with our measurement instruments. In line with our request to fill in at least one log entry per measurement occasion, we set a mean of 1 or more reflections on learning experiences per month as the criterion of affinity. Using the monthly log data to categorize participants into groups thus allowed us to differentiate between participants with regard to the validity of administering logs to them. Moreover, it allowed us to contrast group patterns of dynamics of reflection on the environment, which helps to interpret the results.

For reflection on the environment, we calculated three measures. These measures were based on the daily log data. The first measure was the mean of the intensity of making information explicit in the measurement period per working day (overall information intensity). The mean per working day was calculated to correct for differences between participants in working days.

To assess teachers' routine (or within-person variability) in making information explicit, we applied categorical RQA on each participant's time-series of intensities

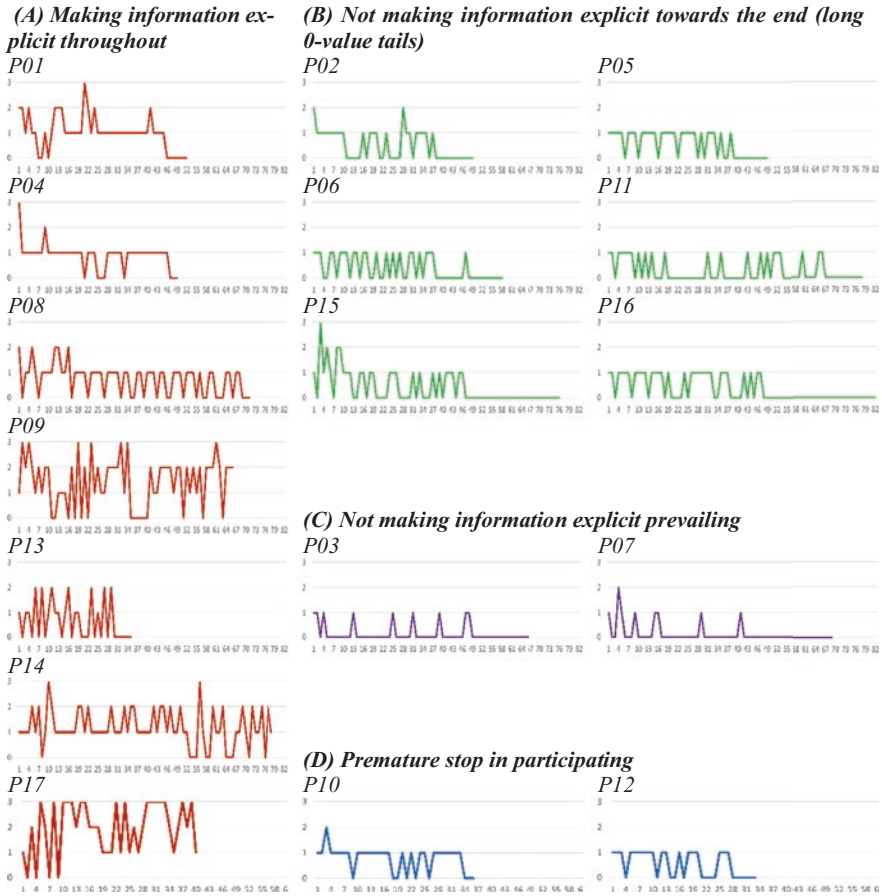


Fig. 11.2 Time-series of participants’ intensities of reflection on the environment
 Note: Reflection intensity = the number of reflection moments per working day. P stands for participant. Numbers indicate the participants. For each graph, time is on the x-axis and reflection intensity is on the y-axis. The time-series only include those days, on which participants received invitations to fill out daily logs (working days). Consequently, the time-series vary in length. The largest number of working days of a participant during the measurement period was 82 and, to ease comparison, this value was set as the length of each x-axis. The time-series have been categorized based on the participants’ response patterns. **(a)**: Mean amount of reflection on learning experiences per month is greater than or equal to 1 ($m_{insights} \geq 1$); **(b–d)**: Mean amount of learning experiences per month is less than 1 ($m_{insights} < 1$). The participants categorized in **(a)** made information explicit using the measurement instrument throughout the measurement period. The participants categorized in **(b)** did not make information explicit using the measurement instrument towards the end of the measurement period (time-series with long 0-value tails), those in **(c)** had time-series in which 0 (no information made explicit using the measurement instrument on a day) prevailed, and those in **(d)** stopped participating prematurely. Consequently, the participants categorized in **(a)** are considered to have more affinity for our measurement instruments, whereas the participants categorized in **(b, c and d)** are considered to have less affinity for them. See Table 11.1 for participants’ measures in each group

Table 11.1 Descriptives and measures of each participant

Participants	Participation capacity		Daily log measures					Monthly log measures		
	FTE	t_{weeks}	t_{days}	Σ_{infos}	m_{infos}	%REC	%DET	t_{months}	$\Sigma_{insights}$	$m_{insights}$
(a) Making information explicit throughout										
01	0.6	18	52(4)	53	1.02	44	76	5(0)	5	1.00
04	0.6	18	49(7)	44	0.90	65	95	5(0)	9	1.80
08	0.8	18	71(3)	54	0.76	47	71	5(0)	7	1.40
09	0.8	18	66(3)	94	1.42	30	47	5(0)	8	1.60
13	0.4	18	35(1)	24	0.69	37	50	5(0)	8	1.60
14	1.0	17	78(4)	92	1.18	40	67	5(0)	11	2.20
17	0.4	18	40(3)	79	1.98	28	42	5(0)	12	2.40
(b) Not making information explicit towards the end (long 0-value tails)										
02	0.6	18	49(5)	25	0.51	46	78	5(0)	0	0.00
05	0.6	18	49(7)	29	0.59	51	74	5(0)	3	0.60
06	0.8	18	58(11)	23	0.40	51	73	5(0)	3	0.60
11	1.0	18	78(13)	22	0.28	59	81	5(0)	0	0.00
15	1.0	17	76(5)	31	0.41	51	75	4(1)	0	0.00
16	1.0	18	82(5)	30	0.37	53	78	5(0)	2	0.40
(c) Not making information explicit prevailing										
03	0.8	18	66(6)	9	0.14	76	94	5(0)	0	0.00
07	0.8	18	69(4)	9	0.13	79	98	5(0)	2	0.40
(d) Premature stop in participating										
10	1.0	11	36(13)	28	0.78	57	80	3(0)	2	0.67
12	1.0	7	33(2)	19	0.58	49	76	n(0)	0	0.00

Note: FTE = Full-Time Equivalent. Here it stands for the number of days per week a participant is employed by the VET college. 1.0 represents an employment of 5 days per week. t_{weeks} = the number of weeks that the participants participated. The measurement period was 18 weeks. Two participants started 1 week later and 2 participants stopped prematurely. t_{days} = the number of working days (i.e. days on which daily log invitations were sent). The value between parentheses is the amount of invitations whose sending had failed. Σ_{infos} = the overall intensity of making information explicit (i.e. the total number of moments of reflection on the environment in the period). Participants could fill in a maximum of 3 daily log entries per working day. m_{infos} = the mean intensity of making information explicit per working day. This measure was calculated to correct for differences between participants in working days and the duration that they participated. %REC = the Recurrence Rate of daily intensities of making information explicit (i.e. recurrences of the number of reflection moments per working day) during the measurement period (as a percentage). %DET = the Determinism of daily intensities of making information explicit (i.e. the number of reflection moments per working day that recur periodically) in the measurement period (as a percentage). t_{months} = the number of months on which monthly log invitations were sent. The value between parentheses is the amount of invitations whose sending had failed. $\Sigma_{insights}$ = overall intensity of making insights explicit (i.e. the total number of moments of reflection on learning experiences in the period). Participants could fill in a maximum of 3 monthly log entries per month (maximum is 15). $m_{insights}$ = the mean intensity of making insights explicit per month. This measure was calculated to correct for differences between participants in the duration that they participated. The descriptives of the participants have been categorized by their response-patterns. **(a)**: $m_{insights} \geq 1$; **(b, c and d)**: $m_{insights} < 1$. Additionally, the participants categorized in **(a)** made information explicit using the measurement instrument throughout the measurement period. The participants categorized in **(b)** did not make information explicit using the measurement instrument towards the end of the measurement period (time-series with long 0-value tails), those in **(c)** had time-series in which 0 (no information made explicit using the measurement instrument on a day) prevailed, and those in **(d)** stopped participating prematurely. See Fig. 11.2 for graphical representations of the participants' time-series

of reflections on the environment per day. The time-series only include those days, on which participants received invitations to fill in daily logs (working days). Other days, such as weekends or holidays, or days of the week on which participants were not employed or were employed by another employer, are not part of the time-series. These ‘non-working days’ were cut out to create an uninterrupted time-series. Consequently, the time-series vary in length. The categorical RQA was conducted in MATLAB, using Marwan’s toolbox (Marwan et al., 2007; Marwan, Wessel, Meyerfeldt, Schirdewan, & Kurths, 2002). As measures of routine, we used Recurrence Rate as a measure of the overall regularity of the intensity of the reflection process over time, and Determinism as a measure of teachers’ persistence in sequences of intensities of reflection. The relations between these four variables were established through visual inspection of scatterplots.

11.4 Results

First, we calculated each measure for each participant. To give an idea of how the trajectories of the intensity of making information explicit (information intensity) correspond to their auto-recurrence plots and their measures, four examples thereof are given in Fig. 11.3.

Second, we assessed the participants’ affinity for the measurement instruments. Seven participants had an overall insight intensity (mean of reflections on learning experiences per month) that was greater than or equal to 1, and thus showed more affinity for the measurement instruments. The other ten participants had an overall insight intensity that was less than 1, and thus showed less affinity for the measurement instruments. Splitting the sample into two groups based on overall insight intensity uncovered striking differences in the temporal patterns of making information explicit. Consider Fig. 11.2. The participants categorized in **(a)** made information explicit using the measurement instrument throughout the measurement period, whereas that seems to falter or cease with the participants in **(b, c, and d)**. The participants categorized in **(b)** did not make information explicit using the measurement instrument towards the end of the measurement period (time-series with long 0-value tails), those in **(c)** had time-series in which 0 (no information made explicit using the measurement instrument on a day) prevailed, and those in **(d)** stopped participating prematurely. Consequently, the participants categorized in **(a)** are considered to have, for whatever reason, more affinity for our measurement instruments in the measurement period, whereas the participants categorized in **(b, c, and d)** are considered to have less affinity for them. Due to the difference between the groups in the fit of the measurement instruments to the participants, administering daily and monthly logs seems to be more valid for the participants in **(a)** than for the others. See Table 11.1 for the participants’ measures and descriptives in each group. A comparison of the descriptives of the two groups suggests a connection between affinity for the measurement instruments and the amount of working days and/or the amount of invitations that failed to be sent.

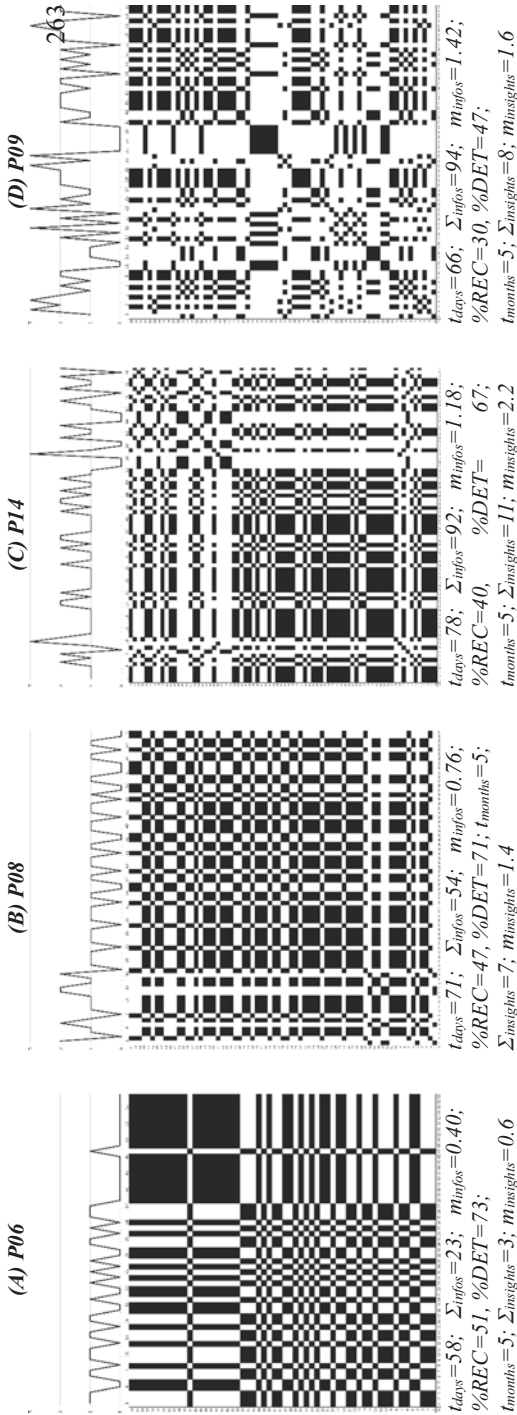


Fig. 11.3 Four examples trajectories of intensities of reflection on the environment represented by time-series and recurrence plots

Note: (a, b, c and d) are the trajectories of intensities of reflection on the environment of 4 participants (based on daily log data). Each trajectory is represented by a time-series graph and a recurrence plot (top and bottom image, respectively). For descriptions of the time-series, see Fig. 11.2 and Table 11.1. t_{days} = the amount of working days during the measurement period. Σ_{infos} = the sum of intensity of making information explicit in the measurement period. m_{infos} = the mean intensity of making information explicit per working day (overall information intensity). $\% \text{REC}$ = the Recurrence Rate of intensities of making information explicit (as a percentage). $\% \text{DET}$ = the Determinism of intensities of making information explicit (as a percentage). For detailed descriptions of the the Recurrence Rate and Determinism, see Fig. 11.1 and the Recurrence Quantification Analysis-section. Additionally, the three measures that are based on the monthly logs are also presented: t_{months} = the number of months on which monthly log invitations were sent. Σ_{insights} = sum of intensity of making insights explicit in the measurement period. m_{insights} = the mean amount of making insights explicit per month (overall insight intensity). For the extent to which overall insights intensity, overall information intensity, Recurrence Rate, and Determinism correlate, see Fig. 11.4 and the Results-section

Third, we explored how overall insight intensity related to the overall information intensity (mean of reflections on the environment per day), and how both related to the Recurrence Rate of information intensity, and Determinism of information intensity. Consider Fig. 11.4. Figure 11.4 Plot (a) suggests a positive correlation between overall information intensity and overall insight intensity within the whole sample, and also within each affinity group separately. More moments of making information explicit co-occurred with more moments of making insights explicit.

Figure 11.4 Plot (b) suggests a negative correlation between overall information intensity and Recurrence Rate within the sample, and also within each affinity group separately. More moments of making information explicit co-occurred with less regularity in doing that. This relation might be explained by the increasing difficulty of having an additional reflection moment beyond the previous one on any given day. Note that none of the participants had both a high level of overall information

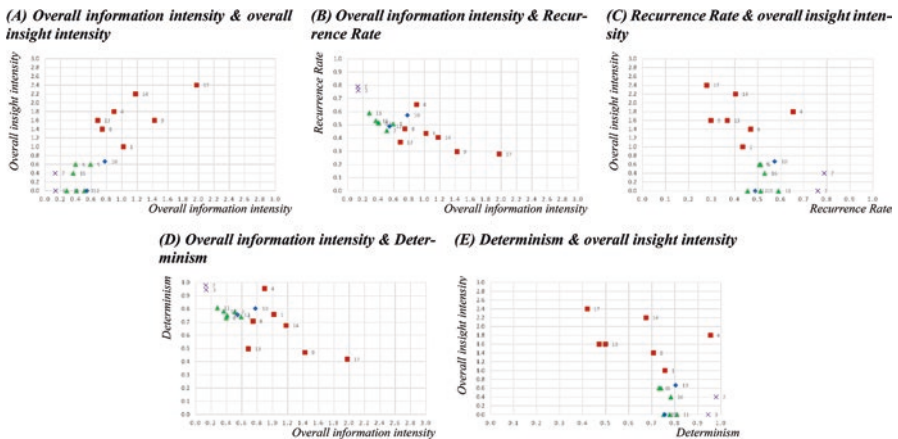


Fig. 11.4 Scatterplots with correlations between overall insight intensity, overall information intensity, Recurrence Rate, and Determinism

Note: Squares represent the group of participants that had more affinity for the measurement instruments (see Fig. 11.2 and Table 11.1). Diamonds, triangles, and crosses represent the group of participants that had less affinity for the measurement instruments. Diamonds represent participants that stopped participating prematurely. Triangles represent participants that did not make information explicit using the measurement instrument towards the end of the measurement period (time-series with long 0-value tails). Crosses represent participants, in whose time-series 0 (no information made explicit using the measurement instrument on a day) prevailed. Numbers indicate the participants. Overall insight intensity = the mean amount of making insights explicit (reflection on learning experiences) per month, overall information intensity = the mean amount of making information explicit (reflection on the environment) per day, Recurrence Rate = Recurrence Rate of information intensities, Determinism = Determinism of information intensities. The means of overall insight intensity (per month) and overall information intensity (per day) for each participant are used to correct for differences between participants in working days and the duration that participants participated. As such, the axis-scales of these two variables go from the minimum (0) to the maximum (3) per measurement occasion. See the text of the Results-section for descriptions of the correlations

intensity and a high Recurrence Rate: a highly regular high level of information intensity did not occur

Figure 11.4 Plot (c) suggests a negative correlation between Recurrence Rate and overall insight intensity in the sample. However, within each affinity group separately, there is no clear relation between Recurrence Rate and overall insight intensity. The group of participants that had more affinity for the measurement instruments made more insights explicit and had less regularity in information intensity during the measurement period than the group of participants that had less affinity for the measurement instruments. The level of making insights explicit seems unrelated to the level of regularity of making information explicit when taking affinity for the measurement instruments into account

Figure 11.4 Plot (d) suggests a negative correlation between overall information intensity and Determinism within the sample, and also within the group of participants that showed more affinity for the measurement instruments. However, within the group of participants that had less affinity for the measurement instruments, there is no clear relation between overall information intensity and Determinism. Note that in this group, nearly all of the information intensity values were 0 or 1 and that both a re-occurrence of 0 as of 1 creates a recurrence point. Due to this small set of low values, this group of participants had a low level of overall information intensity and a high level of Determinism, which was similar for those participants, whose time-series consisted of more 0's as for those, whose time-series consisted of more 1's. For the group of participants that had more affinity for the measurement instruments, more moments of making information explicit co-occurred with less persistent (periodically recurring) engagement in any of the levels of intensity of making information explicit (or sequences thereof). However, this relation can be explained by the difficulty of maintaining a high level of information intensity over time. Indeed, a highly persistent high level of information intensity did not occur. The correlations from both groups thus highlight the weaknesses of using the response rates of daily logs with several entries as a measurement instrument for the application of RQA

Figure 11.4 Plot (e) suggests a negative correlation between Determinism and overall insight intensity in the sample, and also within the group of participants that had more affinity for the measurement instruments. For this group, more moments of making insights explicit co-occurred with less persistent engagement in any of the levels of intensity of making information explicit (or sequences thereof). However, within the group of participants that had less affinity for the measurement instrument, there is no clear relation between Determinism and overall insight intensity. For this group, the level of making insights explicit seems unrelated to the level of persistence of engagement in any of the levels of intensity of making information explicit (or sequences thereof). Following the argumentation given for the relations in plot (d), it seems likely that those participants that manage to make information explicit whenever an opportunity occurs are also the ones that are able to make the most insights explicit. Note that, on the one hand, P04 seems to have organized these opportunities as one per day and thereby to be able to make insights explicit, as inferred from a highly persistent moderate level of information intensity

as well as a high level of overall insight intensity. On the other hand, P17 seems to have strived to have as many of these opportunities as possible on each day and thereby to be able to make insights explicit, as inferred from a lowly persistent high level of information intensity as well as a high level of overall insight intensity

In sum, these results point towards a trend that higher levels of overall information intensity and overall insight intensity concur within a certain period of time. On top of that, no clear pattern was found relating the level of overall insight intensity to with which routine participants made information explicit during a certain period of time.

11.5 Discussion

To summarize, in this study, we explored teacher learning through reflection as a situated and dynamic process using logs as the measurement instruments and RQA as the analysis technique. More specifically, the study focussed on the routine with which teachers engage in making information explicit (reflection on the working environment), and how that, in comparison to the overall levels thereof, relates to making new insights explicit (reflection on learning experiences). We also explored the validity of the measurement instruments and measurement intervals for the application of RQA. Seventeen VET teachers filled in daily and monthly logs over a period of 5 months. From the responses to the daily logs, we generated time-series of the intensity of making information explicit (information intensity) for each participant and applied categorical auto-RQA to each time-series. As measures of the routine of information intensity, Recurrence Rate (regularity) and Determinism (persistence) were used. In addition, we calculated a measure for overall information intensity (the mean amount of information intensity per day in the measurement period) and a measure for overall insight intensity (the mean amount of making insights explicit per month in the measurement period). Relations between the four variables were established through inspection of scatterplots. We found that the sample could be divided into two groups: One that had more and one that had less affinity for the measurement instruments. Moreover, inspection of the scatterplots indicated that higher levels of overall information intensity related to higher levels of overall insight intensity. However, the regularity and the persistence of the intensity with which participants made information explicit had no clear relation with the level of overall insight intensity when taking affinity for the measurement instruments into consideration. In this section we will elaborate on these findings.

That the sample could be divided into one group that had more and another group that had less affinity for the measurement instruments (both daily and monthly logs), may be due to several related reasons. One reason might be related to the difference between the groups in the amount of invitations that failed to be sent. The participants in the less affinity group did not receive an invitation about twice as often as the participants in the more affinity group when correcting for the amount of working days. Increasingly, undependability may have led teachers to falter or

cease using our measurement instruments. One of the challenges in conducting this study was to send personalized logs with personalized intervals using an online instrument that was not designed for that, but rather for large-scale, cross-sectional surveys. The developments in digital technology, such as smartphone applications, will have made this problem obsolete for future studies, however.

A second reason might be related to the difference between the groups in the number of days per week they worked. The participants in the less affinity group worked roughly a day more than the participants in the more affinity group, and may simply have been too busy to disengage from the immediacy of their work to make time to reflect by using logs.

A third reason might be related to the dynamics of the reflection process itself. As experience grows, people become less responsive to new information in their environment, and the new information is not further corroborated into experience (Schöner & Dineva, 2007). In this study, the daily logs served as impulses to become aware of information in the environment that some participants might otherwise not have made explicit. Consequently, as experience with this initially attended-to information grew, participants may have felt a need to consolidate acting upon that information first, rather than attending to even more information and deciding how to act upon that. This reason seems particularly fitting for the participants, who did not make information explicit using the measurement instrument towards the end of the measurement period. Nevertheless, whereas administering logs seems to be less valid for these particular participants, the dense time-series the logs generated did point towards an interesting dynamic that future research may explore further.

This third reason relates to that teachers need time to learn (and can attend to teaching less), and also need time to teach (and can attend to learning less) (Mulford, 2010), which points towards the fourth reason: Despite the fact that all teachers volunteered to participate, it could have been that the participants in the more affinity group had a period in which they could attend to learning more, whereas the participants in the less affinity group had a period in which they had to attend to teaching. This fourth reason might complement the second reason.

One final reason may be that the participants did develop and adapt their teaching practices, but not through reflection on the working environment and learning experiences at a later point. Rather, they may have engaged in experimentation with new teaching methods or keeping up to date with the latest literature (Oude Groote Beverborg, Slegers, & van Veen, 2015c). Despite their initial willingness to participate, they may have found that making information and insights explicit by using logs did not benefit them. Future studies could investigate for whom what knowledge content is discovered with what additional learning activities or other forms of reflection. All in all, using daily and monthly logs with open questions to study learning through reflection fitted better to some participants than to others.

For the discussion about the findings related to how the extent to which the overall level and the routine of the intensity of making information explicit co-occurs with the overall intensity of making insights explicit, we focus on the group of participants that was considered to have more affinity for the measurement instruments. We found that levels of overall reflection on the working environment

positively related to levels of overall reflection on learning experiences. In this regard, it is relevant that information to be made explicit is always present in the working environment. Insights, on the contrary, can only be made explicit when learning experiences occurred. As such, the situated manner in which we assessed teacher learning through reflection corroborates findings from large-scale survey studies, which showed that engaging in learning activities more goes together with having more learning results (Oude Groote Beverborg et al., 2015a; Slegers et al., 2014).

Furthermore, we found no clear relation between the measures of the routine with which teachers reflect on the working environment and their overall reflection on learning experiences. The regularity of making information explicit was unrelated to the overall level of making insights explicit. The persistence of making information explicit could be seen as negatively correlated to the overall level of making insights explicit, but the dispersion was high. To illustrate, from the top three participants in making insights explicit, one had the least and one had the most persistence in the intensity of making information explicit. Thus, the answer to the question about whether learning can be facilitated through reflecting very constantly or in bursts, is: both. The application of RQA thereby extends research on sequences of (multiple) learning activities (Endedijk, Hoekman, & Slegers, 2014; Zwart et al., 2008). Moreover, these RQA-based findings suggest that constancy in reflection-intensity is not necessarily beneficial to school improvement and educational change (see also Mulford, 2010; Weick, 1996). Such constancy may, again, fit better to some than to others. Consequently, teachers cannot be discharged from the responsibility of finding out what manner of learning befits them personally, and colleagues can only seduce them to do so. Studies with additional measures and in additional contexts are needed to validate our findings concerning the constancy of everyday teacher learning.

How, then, to support teachers in sustaining levels of reflection without enforcing high constancy thereof (see also Giles & Hargreaves, 2006; Timperley & Alton-Lee, 2008)? An answer thereon may not be based on focussing on the routine of engagement in the learning activity itself, but by also taking the situated nature of the process in consideration (Barab et al., 1999). Our findings suggest that those participants that are able to make the most insights explicit are also the ones that manage to make information explicit whenever an opportunity occurs, which could be done by organizing such opportunities (i.e. moments of disengagement from the work flow, the use of evaluation instruments or logs, classroom observations, meetings, or appraisal interviews) with determined intervals, but also by being keen to have as many such moments as the working environment may provide each day, or a combination of both. Either way, the working environment would have to provide ample information that is salient and interesting enough to further think about and to distil a new way of acting from, whenever teachers have an opportunity to do so (Lohman & Woolf, 2001). In this respect, critically reflecting colleagues and transformational school leaders, who inspire, support, and stimulate, are crucial in helping to see the workplace in a new light and in providing examples of how one can synchronize one's practice with newly found information (Hoekstra & Korthagen, 2011; Oude

Groote Beverborg et al., 2015c; van Woerkom, 2010). Future research could investigate the development and dynamics of coordination of team members in creating such an interesting environment by engaging in knowledge sharing with the aim to co-construct shared meaning and to facilitate school improvement and educational change (see also Zoethout, Wesselink, Runhaar, & Mulder, 2017).

In sum, the findings of this study indicate that teachers who make more information from their working environment explicit are also able to make more new insights explicit. This suggests that higher levels of engagement in reflection are beneficial to teachers' developments, and, by extension, to educational change and school improvement. The routine with which teachers make information explicit was found to be mostly unrelated to making new insights explicit. Of importance seems to be to reflect upon the working environment whenever an opportunity arises. Crucial seems to be that this (social) environment provides information that is salient and interesting enough to distil a new way of acting and attending from. Teachers might additionally benefit sometimes from organizing opportunities to become aware of information in the environment with a certain constancy. In this regard, the use of daily and monthly logs seems to fit better to some participants than to others.

This study is a first step in understanding teacher learning through reflection in the workplace as an everyday and ongoing process. The use of measurement instruments that generate dense time-series and the application of RQA to assess stability and flexibility over time shows that longitudinal research can concentrate on more than just on growth or couplings between variables over time (e.g. Hallinger & Heck, 2011; Heck & Hallinger, 2009; Heck & Hallinger, 2010; Oude Groote Beverborg et al., 2015a; Slegers et al., 2014; Smylie & Wenzel, 2003; Thoonen, Slegers, Oort, Peetsma, & Geijsel, 2011). Moreover, the study provides an example of how novel methodology, such as RQA, can be adopted to tap into professional learning as a dynamic and situated process in support of school improvement and educational change.

11.5.1 Limitations & Future Directions

The initial idea of the study was to dive deeper into the reflection process than presented here, by measuring what specific types of information teachers attended to using the daily logs, by measuring the contents of learning experiences using the monthly logs, by analysing the dynamics of attending to those types of information using categorical auto-RQA, and by establishing a relation between for instance the persistence in one type of information and the occurrence of a learning experience with a corresponding content. With this aim, we coded the daily and monthly log entries. However, the time-series generated per code-category were not dense enough for the application of RQA. Moreover, we assumed that setting a fixed time for reporting learning experiences would help generate a higher response rate. However, not knowing when learning experiences took place during the months

made it very difficult to relate it to the information reported in the daily logs. Thus, the design failed to generate the timing information that would have been needed to be able to model the learning experiences' occurrences. Having participants fill in learning experiences at (or very soon after) the moment they have them, would therefore have been a better approach. Additionally, our choice of measurement interval was a compromise between the expected rate of change with which salient information would be made explicit and the practical consideration of not wanting to burden the participating teachers too much. Our measurement intervals were therefore too crude for our initial purposes. In sum, measurement methods with a higher sampling rate, such as observations that happen in real-time, are needed to model how information in the working environment affords development and adaptation more accurately (Granic & Patterson, 2006; Lewis et al., 1999; Lichtwarck-Aschoff et al., 2012). Nevertheless, qualitative analyses on the data generated by the logs used in this study can be used to relate the contents that teachers reflected upon with the contents of what they learnt. This would still contribute to understanding more about the role of affordances in teacher learning, but the aim would no longer lie on finding systematic relationships (Barab & Roth, 2006; Greeno, 1994; Little, 2003; Maitlis, 2005).

We would like to stress that RQA's derive their power from frequent measurements – and not from a large sample size. Whereas using small samples could constrain generalizability, studies assessing for instance the temporal pattern of teacher interactions in only one team in real-time, might provide important, new insights into the process of how teachers collaborate to make sense of the challenges they face and how that culminates in the generation of new knowledge or a shared meaning (e.g. Fullan, 2007). Additionally, such studies might prove very valuable for researchers, who are interested in the systematics of change processes and seek to combine the results of various studies in simulation studies (Clarke & Hollingsworth, 2002), rather than meta-analyses (see also Richter, Dawson, & West, 2011; Sun & Leithwood, 2012; Witziers, Bosker, & Krüger, 2003). By building on the current study, future research could contribute to a bottom-up understanding of how learning communities, but also change capacity of schools, emerge and continue to evolve (Hopkins, Harris, Stoll, & Mackay, 2010; Stoll, 2009). Another benefit of the proposed measurement methods and analyses, due to their focus on the circumstances and periodicities of individuals, is that they allow for tailored advice to individual teachers (or teams of teachers). Consequently, this approach to investigating professional learning would allow teachers and policy makers alike to formulate situated expectations about the pace of adaptation, the rate of innovations within a certain time, and delays in proficiency. An interesting follow-up question nevertheless concerns the extent to which diaries served as an intervention for fostering reflective learning and thus influenced the learning occurrences accordingly (Hoekstra & Korthagen, 2011). A new study with an experimental design and additional dependent measures would be needed to investigate this (Maag Merki, 2014).

Despite its limitations, this study does provide a first enquiry into studying teacher learning as a situated and dynamic process through the use of logs and RQA. In future research, the methodology could have utility in studying aspects of

the dynamics of teacher learning such as, on an individual level, shifts in appreciation of the importance of certain classroom practices or differentiation in perception, or on an organizational level, alternations of periods of tight versus loose couplings between teachers, teams, or departments (see also Korthagen, 2010; Kunnen & Bosma, 2000; Mulford, 2010; Nonaka, 1994; Orton & Weick, 1990). The methodology could also help policy makers in balancing top-down and bottom-up processes in shaping the organization of the school (e.g. Feldhoff, Huber, & Rolff, 2010; Hopkins et al., 2010; Spillane et al., 2002; van der Vegt & van de Vliert, 2002). Moreover, by studying the temporal pattern of sensemaking processes in schools (see also Coburn, 2001; Feldhoff & Wurster, 2017; Spillane et al., 2002), more can be understood about the development of professional learning communities and the inner workings of the change capacities of schools. Consequently, in line with trends in accountability to focus on learning of organizations rather than fulfilment of inspection criteria, Inspectorates of Education could use the methodology to tap into a developmental process rather than only the results thereof in order to support the sensemaking processes in schools (Feldhoff & Wurster, 2017).

Acknowledgement The authors would like to thank Simone Kühn and Barbara Müller for their invaluable advice.

Appendices

Appendix A

Daily Log 1(2)

Information

This question is about informal learning from colleagues in the workplace.

Informal learning can be seen as the daily discovery of information.

Information can be known or new, it can be positive or negative, and it can be something from the educational praxis or something from a conversation.

More concretely, you can think of information as something a colleague said; something that was recommended to you; something you experienced; the manner in which you did something; the feedback you gave someone; something you did not do; etc.

This question is about which information struck you the most today. Below you see four answer categories.

Below, you see four answer categories.

Choose one of the answer categories.

Later, you can choose a new answer category.

After you have clicked on one of the options, you will be presented with questions about the nature of the information that struck you.

After you have answered the questions about the nature of the information, you can choose one of the four answer categories again.

You can choose an answer category maximally three times, thereafter the diary entry of today will stop.

Try to use no more than 5 min for filling in today's diary entries.

Which of the options below struck you the most today?

(Where "colleague" is stated, you can also read "colleagues")

- I agreed with something a colleague said or did
- I disagreed with something a colleague said or did
- Something a colleague did helped me
- Something a colleague did hindered me

PREVIOUS page

NEXT page

Daily Log 2(2)

Information

Where "colleague" is stated, you can also read "colleagues".

You stated that you agreed with something a colleague said or did today.⁶

The following questions elaborate on that.

Try to answer the open questions in no more than three sentences.

⁶In case another answer category was selected on the previous page, the text throughout this page was adapted accordingly.

What did your colleague say or do today?

What about what your colleague said or did was relevant for you?

(If needed, you can select more than one option, but try to constrain your answer to one option.)

- That a colleague said or did something
- That that colleague said or did something
- What the colleague said or did
- Something about what the colleague said or did (e.g., that one sentence or action)
- The result of what the colleague said or did
- All of the colleagues performance

Otherwise, namely...

What was the task that you worked on, to which what your colleague said or did related?

What was your reaction to what your colleague said or did?

To what extent did you agree with what your colleague said or did?

- 1: I agreed a little
- 2: I agreed
- 3: I agreed a lot

Do you intend to attend to it in the following weeks?

- Yes
- No
- Does not apply

PREVIOUS page

NEXT page

Appendix B

Monthly Log

Learning Experience

Learning can occur everywhere and always. Learning can be planned and spontaneous. You become conscious of having learned something when you have had a learning experience.

You can think for example of a learning experience as having found a new way to prepare a task with your colleagues, or as having had an insight about how you can transfer something to your students after having had a conversation with a colleague.

The questions in the monthly log are about learning experiences that you have had in the past month. We kindly ask you to report three learning experiences.⁷ Each entry is about one learning experience. This is the entry of learning experience 1⁸.

Try to answer the questions in no more than three sentences.

⁷Although we kindly asked to report three learning experiences, it was voluntary whether participants filled in 0, 1, 2, or 3 monthly log entries.

⁸For the second and third entry filled in within the log of 1 month, this number is 2 or 3, respectively.

1. What did you learn in the past month?

2. For the performance of which task, was what was learned relevant?

3. To which personal or professional development goal did what was learned relate?

4. What was needed to learn it?

(Think for instance of what knowledge, skills, experiences, means, or people)

5. In which way did you learn it?

6. Why do you learn it in this specific way?

7. How did you find out that you had learned something?

Describe the learning experience.

(i.a. with whom, working on which task, etc.)

8. With which aspects of the learning process are you satisfied, and what would you do differently next time?

9. Now that you have learned this, what will you attend to in the following weeks?

10. On the basis of this learning experience, which personal or professional goal do you set for yourself for the following weeks?

PREVIOUS page

NEXT page

References

- Argyris, C., & Schön, D. A. (1974). *Theory in practice: Increasing professional effectiveness*. San Francisco, CA: Jossey-Bass.
- Barab, S. A., Cherkas-Julkowski, M., Swenson, R., Garrett, S., Shaw, R. E., & Young, M. (1999). Principles of self-organization: Learning as participation in autocatakinetic systems. *Journal of the Learning Sciences*, 8(3–4), 349–390.
- Barab, S. A., & Roth, W. M. (2006). Curriculum-based ecosystems: Supporting knowing from an ecological perspective. *Educational Researcher*, 35(5), 3–13.
- Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford Press.
- Brick, T. R., Gray, A. L., & Staples, A. D. (2018). Recurrence quantification for the analysis of coupled processes in aging. *The Journals of Gerontology: Series B*, 73(1), 134–147.
- Broer, H., & Takens, F. (2009). *Dynamical systems and chaos*. Utrecht, The Netherlands: Epsilon Uitgaven.
- Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, 18, 947–967.
- Coburn, C. E. (2001). Collective Sensemaking about Reading: How teachers mediate Reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, 23(2), 145–170.
- Coburn, C. E. (2004). Beyond decoupling: Rethinking the relationship between the institutional environment and the classroom. *Sociology of Education*, 77, 211–244.
- Coburn, C. E. (2005). Shaping teacher sensemaking: School leaders and the enactment of reading policy. *Educational Policy*, 19(3), 476–509.
- Coburn, C. E. (2006). Framing the problem of reading instruction: Using frame analysis to uncover the microprocesses of policy implementation. *American Educational Research Journal*, 43(3), 343–349.
- Cox, R. F. A., Van der Steen, S., Guevara Guerrero, M., Hoekstra, L., & Van Dijk, M. (2016). Chromatic and anisotropic cross-recurrence quantification analysis of interpersonal behavior. In C. Webber, C. Ioana, & N. Marwan (Eds.), *Recurrence Plots and Their Quantifications: Expanding Horizons: Proceedings of the 6th International Symposium on Recurrence Plots*, Grenoble, France, 17–19 June 2015. (Springer Proceedings in Physics). Springer.
- Dale, R., & Spivey, M. J. (2005). Categorical recurrence analysis of child language. In *Proceedings of the 27th annual meeting of the cognitive science society* (pp. 530–535). Mahwah, NJ: Lawrence Erlbaum.
- de Groot, E., Endedijk, M. D., Jaarsma, A. D. C., Simons, P. R. J., & van Breukelen, P. (2014). Critically reflective dialogues in learning communities of professionals. *Studies in Continuing Education*, 36(1), 15–37.
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38, 181–199.
- Endedijk, M. D. (2010). *Student teachers' self-regulated learning*. (Doctoral dissertation). Utrecht, The Netherlands: University Utrecht.
- Endedijk, M. D., Brekelmans, M., Verloop, N., Slegers, P. J. C., & Vermunt, J. D. (2014). Individual differences in student teachers' self-regulated learning: An examination of regulation configurations in relation to conceptions of learning to teach. *Learning and Individual Differences*, 30, 155–162.
- Endedijk, M. D., Hoekman, M., & Slegers, P. J. C. (2014). Learning paths of engineers: Studying sequences of learning activities to understand knowledge workers' professional development. In *7th EARLI SIG 14 Conference*. Oslo, Norway, 27–29 August 2014b.
- Feldhoff, T., Huber, S. G., & Rolff, H. G. (2010). Steering groups as designers of school development processes. *Journal for Educational Research Online*, 2(2), 98–124.
- Feldhoff, T., Radisch, F., & Bischof, L. M. (2016). Designs and methods in school improvement research: A systematic review. *Journal of Educational Administration*, 54(2), 209–240.

- Feldhoff, T., Radisch, F., & Klieme, E. (2014). Methods in longitudinal school improvement: State of the art. *Journal of Educational Administration*, 52(5).
- Feldhoff, T., & Wurster, S. (2017). Ein Angebot das sie nicht ablehnen können? Zur Funktion von Schulinspektionsergebnissen als Deutungsangebot zur Unterstützung schulischer Verarbeitungsprozesse und schulische Reaktionsweisen auf das Angebot. *Empirische Pädagogik*.
- Fullan, M. (2007). *The new meaning of educational change* (4th ed.). London, UK: Teachers College Press.
- Geursen, J., de Heer, A., Korthagen, F. A., Lunenberg, M., & Zwart, R. (2010). The importance of being aware: Developing professional identities in educators and researchers. *Studying Teacher Education*, 6(3), 291–302.
- Gibson, J. J. (1979/1986). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Giles, C., & Hargreaves, A. (2006). The sustainability of innovative schools as learning organizations and professional learning communities during standardized reform. *Educational Administration Quarterly*, 42(1), 124–156.
- Granic, I., & Dishion, T. J. (2003). Deviant talk in adolescent friendships: A step toward measuring a pathogenic attractor process. *Social Development*, 12(3), 314–334.
- Granic, I., & Patterson, G. R. (2006). Toward a comprehensive model of antisocial development: A dynamic systems approach. *Psychological Review*, 113(1), 101–131.
- Greeno, J. G. (1994). Gibson's affordances. *Psychological Review*, 101(2), 336–342.
- Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist*, 53(1), 5–26.
- Guastello, S. J. (2002). *Managing emergent phenomena. Nonlinear dynamics in work organizations*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Guastello, S. J., Johnson, E. A., & Rieke, M. L. (1999). Nonlinear dynamics of motivational flow. *Nonlinear Dynamics, Psychology, and Life Sciences*, 3, 259–273.
- Hallinger, P., & Heck, R. (2011). Exploring the journey of school improvement: Classifying and analyzing patterns of change in school improvement processes and learning outcomes. *School Effectiveness and School Improvement*, 22, 149–173.
- Hamaker, E. L. (2012). Why researchers should think “within-person”: A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford.
- Heck, R. H., & Hallinger, P. (2009). Assessing the contribution of distributed leadership to school improvement and growth in math achievement. *American Educational Research Journal*, 46, 626–658.
- Heck, R. H., & Hallinger, P. (2010). Collaborative leadership effects on school improvement: Integrating unidirectional- and reciprocal-effects models. *The Elementary School Journal*, 111(2), 226–252.
- Heft, H. (2001). *Ecological psychology in context: James Gibson Roger Barker and the legacy of William James's radical empiricism*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Hoekstra, A., & Korthagen, F. (2011). Teacher learning in a context of educational change: Informal learning versus systematically supported learning. *Journal of Teacher Education*, 62(1), 76–92.
- Hopkins, D., Harris, A., Stoll, L., & Mackay, T. (2010). School and system improvement: State of the art review. Keynote presentation prepared for the 24th. *International Congress of School Effectiveness and School Improvement*.
- Horn, I. S. (2005). Learning on the job: A situated account of teacher learning in high school mathematics departments. *Cognition and Instruction*, 23(2), 207–236.
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
- Korthagen, F., & Vasalos, A. (2005). Levels in reflection: Core reflection as a means to enhance professional growth. *Teachers and Teaching: Theory and Practice*, 11(1), 47–71.
- Korthagen, F. A. J. (2001). *Linking practice and theory: The pedagogy of realistic teacher education*. Paper presented at the Annual Meeting of the American Educational Research Association. Seattle, WA.

- Korthagen, F. A. J. (2010). Situated learning theory and the pedagogy of teacher education: Towards an integrative view of teacher behaviour and teacher learning. *Teacher and Teacher Education*, 26(1), 98–106.
- Kugler, P. N., Shaw, R. E., Vicente, K. J., & Kinsella-Shaw, J. (1990). Inquiry into intentional systems I: Issues in ecological physics. *Psychological Research*, 52(2–3), 98–121.
- Kunnen, E. S., & Bosma, H. A. (2000). Development of meaning making: A dynamic systems approach. *New Ideas in Psychology*, 18(1), 57–82.
- Lave, J., & Wenger, E. (1991). *Situated learning. Legitimate peripheral participation*. New York, NY: Cambridge University Press.
- Lewis, M. D., Lamey, A. V., & Douglas, L. (1999). A new dynamic systems method for the analysis of early socioemotional development. *Developmental Science*, 2(4), 457–475.
- Lichtwarck-Aschoff, A., Hasselman, F., Cox, R., Pepler, D., & Granic, I. (2012). A characteristic destabilization profile in parent-child interactions associated with treatment efficacy for aggressive children. *Nonlinear Dynamics, Psychology, and Life Sciences*, 16(3), 353–379.
- Lichtwarck-Aschoff, A., Kunnen, S. E., & van Geert, P. L. (2009). Here we go again: A dynamic systems perspective on emotional rigidity across parent-adolescent conflicts. *Developmental Psychology*, 45(5), 1364–1375.
- Little, J. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *The Teachers College Record*, 91(4), 509–536.
- Little, J. W. (2003). Inside teacher community: Representations of classroom practice. *Teachers College Record*, 105(6), 913–945.
- Little, J. W., & Horn, I. S. (2007). Normalizing' problems of practice: Converting routine conversation into a resource for learning in professional communities. In L. Stoll & K. S. Louis (Eds.), *Professional learning communities: Divergence depth and dilemmas* (pp. 79–92). London, UK: McGraw-Hill Education (UK).
- Lohman, M. C., & Woolf, N. H. (2001). Self-initiated learning activities of experienced public school teachers: Methods, sources, and relevant organizational influences. *Teachers and Teaching*, 7, 59–74.
- Lunenburg, M., Korthagen, F., & Zwart, R. (2011). Self-study research and the development of teacher educator's professional identities. *European Educational Research Journal*, 10(3), 407–420.
- Lunenburg, M., Zwart, R., & Korthagen, F. (2010). Critical issues in supporting self-study. *Teaching and Teacher Education*, 26(6), 1280–1289.
- Maag Merki, K. (2014). Conducting intervention studies on school improvement: An analysis of possibilities and constraints based on an intervention study of teacher cooperation. *Journal of Educational Administration*, 52(5), 590–616.
- Maag Merki, K., Grob, U., Rechsteiner, B., Rickenbacher, A., & Wullschleger, A. (2021). Regulation activities of teachers in secondary schools. Development of a theoretical framework and exploratory analyses in four secondary schools based on time sampling data. In A. Oude Groote Beverborg, K. Maag Merki, F. Feldhoff, & F. Radisch (Eds.), *Concept and design developments in school improvement research. State of the art longitudinal, multilevel, and mixed methods and their relevance for educational accountability* (pp. 257–301). Dordrecht, The Netherlands: Springer.
- Mainhard, M. T., Pennings, H. J., Wubbels, T., & Brekelmans, M. (2012). Mapping control and affiliation in teacher-student interaction with State Space Grids. *Teaching and Teacher Education*, 28(7), 1027–1037.
- Maitlis, S. (2005). The social processes of organizational sensemaking. *Academy of Management Journal*, 48(1), 21–49.
- Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5–6), 237–329.
- Marwan, N., Wessel, N., Meyerfeldt, U., Schirdewan, A., & Kurths, J. (2002). Recurrence plot based measures of complexity and its application to heart rate variability data. *Physical Review E*, 66(2), 26702.

- Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, 18(2), 112–117.
- Mulford, B. (2010). Recent developments in the field of educational leadership: The challenge of complexity. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *Second international handbook of educational change* (pp. 187–208). Dordrecht, The Netherlands: Springer.
- Ng, F. S. D. (2021). Reframing educational leadership research in the 21st century. In A. Oude Groote Beverborg, K. Maag Merki, F. Feldhoff, & F. Radisch (Eds.), *Concept and design developments in school improvement research. State of the art longitudinal, multilevel, and mixed methods and their relevance for educational accountability* (pp. 107–135). Dordrecht, The Netherlands: Springer.
- Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science*, 5(1), 14–37.
- O'Brien, B. A., Wallot, S., Haussmann, A., & Kloos, H. (2014). Using complexity metrics to assess silent reading fluency: A cross-sectional study comparing oral and silent reading. *Scientific Studies of Reading*, 18(4), 235–254.
- Orton, J. D., & Weick, K. E. (1990). Loosely coupled systems: A reconceptualization. *Academy of Management Review*, 15(2), 203–223.
- Oude Groote Beverborg, A. (2015). *Fostering sustained teacher learning: Co-creating purposeful and empowering workplaces*. (Doctoral dissertation). Enschede, The Netherlands: University of Twente.
- Oude Groote Beverborg, A., Slegers, P. J. C., Endedijk, M. D., & van Veen, K. (2015a). Towards sustaining levels of reflective learning: how do transformational leadership, task interdependence, and self-efficacy shape teacher learning in schools? *Societies*, 5, 187–219.
- Oude Groote Beverborg, A., Slegers, P. J. C., & van Veen, K. (2015b). Fostering teacher learning in VET colleges: Do leadership and teamwork matter? *Teaching and Teacher Education*, 48, 22–33.
- Oude Groote Beverborg, A., Slegers, P.J.C., & van Veen, K. (2015c). Promoting VET teachers' individual and social learning activities: the empowering and purposeful role of transformational leadership, interdependence, and self-efficacy. *Empirical Research in Vocational Education and Training*, 7(5).
- Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, 29(1), 4–15.
- Reed, E. S. (1996). *Encountering the world. Toward an ecological psychology*. New York, NY: Oxford University Press.
- Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science*, 18(5), 407–413.
- Richardson, M. J., Schmidt, R. C., & Kay, B. A. (2007). Distinguishing the noise and attractor strength of coordinated limb movements using recurrence analysis. *Biological Cybernetics*, 96(1), 59–78.
- Richter, A. W., Dawson, J. F., & West, M. A. (2011). The effectiveness of teams in organizations: A meta-analysis. *The International Journal of Human Resource Management*, 22(13), 2749–2769.
- Riley, M. A., & Van Orden, G. C. (2005). *Tutorials in contemporary nonlinear methods for the behavioral sciences*. Retrieved March 1, 2005, from <http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp>
- Schön, D. (1983). *The reflective practitioner: How professionals think in action*. New York, NY: Basic Books.
- Schöner, G., & Dineva, E. (2007). Dynamic instabilities as mechanisms for emergence. *Developmental Science*, 10(1), 69–74.
- Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 326.

- Sleegers, P. J. C., & Spillane, J. P. (2009). In pursuit of school leadership and management expertise: Introduction to the special issue. *Leadership and Policy in Schools*, 8, 121–127.
- Sleegers, P. J. C., Thoonen, E. E., Oort, F. J., & Peetsma, T. T. (2014). Changing classroom practices: The role of school-wide capacity for sustainable improvement. *Journal of Educational Administration*, 52(5), 617–652.
- Sleegers, P. J. C., Wassink, H., van Veen, K., & Imants, J. (2009). School leaders' problem framing: A sense-making approach to problem-solving processes of beginning school leaders. *Leadership and Policy in Schools*, 8(2), 152–172.
- Smylie, M. A., & Wenzel, S. A. (2003). The Chicago Annenberg challenge: Successes, failures, and lessons for the future. *Final Technical Report of the Chicago Annenberg Research Project*.
- Spillane, J. P., & Miele, D. B. (2007). Evidence in practice: A framing of the terrain. In P. A. Moss (Ed.), *Evidence and decision-making: The 106th yearbook of the National Society for the Study of Education, Part I* (pp. 46–73). Malden, MA: Blackwell Publishing.
- Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocussing implementation research. *Review of Educational Research*, 72(3), 387–431.
- Spillane, J. P., & Zuberi, A. (2021). Designing and piloting a leadership daily practice log: Using logs to study the practice of leadership. In A. Oude Groote Beverborg, K. Maag Merki, F. Feldhoff, & F. Radisch (Eds.), *Concept and design developments in school improvement research. State of the art longitudinal, multilevel, and mixed methods and their relevance for educational accountability* (pp. 155–195). Dordrecht, The Netherlands: Springer.
- Staples, D. S., & Webster, J. (2008). Exploring the effects of trust, task interdependence and virtualness on knowledge sharing in teams. *Information Systems Journal*, 18(6), 617–640.
- Steenbeek, H. W., & van Geert, P. L. C. (2007). A theory and dynamic model of dyadic interaction: Concerns, appraisals, and contagiousness in a developmental context. *Developmental Review*, 27, 1–40.
- Stephen, D. G., & Dixon, J. A. (2009). The self-organization of insight: Entropy and power laws in problem solving. *The Journal of Problem Solving*, 2(1), 72–101.
- Stephen, D. G., Dixon, J. A., & Isenhower, R. W. (2009). Dynamics of representational change: Entropy, action, and cognition. *Journal of Experimental Psychology. Human Perception and Performance*, 35(6), 1811–1832.
- Stoll, L. (2009). Capacity building for school improvement or creating capacity for learning? A changing landscape. *Journal of Educational Change*, 10, 115–127.
- Sun, J., & Leithwood, K. (2012). Transformational school leadership effects on student achievement. *Leadership and Policy in Schools*, 11(4), 418–451.
- Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980* (pp. 366–381). Heidelberg, Germany: Springer.
- Thoonen, E. E. J., Sleegers, P. J. C., Oort, F. J., Peetsma, T. T. D., & Geijsel, F. P. (2011). How to improve teaching practices: The role of teacher motivation, organizational factors, and leadership practices. *Educational Administration Quarterly*, 47(3), 496–536.
- Timperley, H., & Alton-Lee, A. (2008). Reframing teacher professional learning: An alternative policy approach to strengthening valued outcomes for diverse learners. *Review of Research in Education*, 32(1), 328–369.
- van der Lans, R. M. (2018). On the “association between two things”: The case of student surveys and classroom observations of teaching quality. *Educational Assessment, Evaluation and Accountability*, 30(4), 347–366.
- van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education*, 86(2), 247–264.
- van der Vegt, G., & van de Vliert, E. (2002). Intragroup interdependence and affectiveness: Review and proposed directions for theory and practice. *Journal of Managerial Psychology*, 17(1), 50–67.
- van Geert, P., & Steenbeek, H. (2005). Explaining after by before: Basic aspects of a dynamic systems approach to the study of development. *Developmental Review*, 25, 408–442.

- van Geert, P., & Steenbeek, H. (2014). The good, the bad and the ugly? The dynamic interplay between educational practice, policy and research. *Complicity: An International Journal of Complexity and Education*, 11(2), 22–39.
- van Woerkom, M. (2004). The concept of critical reflection and its implications for human resource development. *Advances in Developing Human Resources*, 6(2), 178–192.
- van Woerkom, M. (2010). Critical reflection as a rationalistic ideal. *Adult Education Quarterly*, 60(4), 339–356.
- Voestermans, P., & Verheggen, T. (2007). *Cultuur en lichaam: Een cultuurpsychologisch perspectief op patronen in gedrag*. Oxford, UK: Blackwell.
- Voestermans, P., & Verheggen, T. (2013). *Culture as embodiment: The social tuning of behavior*. Chichester, UK: John Wiley Sons.
- Weick, K. E. (1996). Enactment and the boundaryless career: Organizing as we work. In M. B. Arthur & D. M. Rousseau (Eds.), *The boundaryless career: A new employment principle for a new organizational era* (pp. 40–57). New York, NY: Oxford University Press Inc.
- Weick, K. E. (2006). The role of imagination in the organizing of knowledge. *European Journal of Information Systems*, 15, 446–452.
- Weick, K. E. (2011). Organized sensemaking: A commentary on processes of interpretive work. *Human Relations*, 65(1), 141–153.
- Wenger, E. (1998). *Communities of practice: Learning meaning and identity*. New York, NY: Cambridge University Press.
- Wijnants, M. L., Bosman, A. M., Hasselman, F. W., Cox, R. F., & Van Orden, G. C. (2009). 1/f scaling in movement time changes with practice in precision. *Nonlinear Dynamics, Psychology, and Life Sciences*, 13(1), 75–94.
- Wijnants, M. L., Hasselman, F., Cox, R. F. A., Bosman, A. M. T., & Van Orden, G. (2012). An interaction-dominant perspective on reading fluency and dyslexia. *Annals of Dyslexia*, 62(2), 100–119.
- Witziers, B., Bosker, R. J., & Krüger, M. L. (2003). Educational leadership and student achievement: The elusive search for an association. *Educational Administration Quarterly*, 39(3), 398–425.
- Zoethout, H., Wesseling, R., Runhaar, P., & Mulder, M. (2017). Using Transactivity to understand emergence of team learning. *Small Group Research*, 48(2), 190–214.
- Zwart, R. C., Wubbels, T., Bergen, T. C. M., & Bolhuis, S. (2007). Experienced teacher learning within the context of reciprocal peer coaching. *Teachers and Teaching: Theory and Practice*, 13(2), 165–187.
- Zwart, R. C., Wubbels, T., Bolhuis, S., & Bergen, T. C. M. (2008). Teacher learning through reciprocal peer coaching: An analysis of activity sequences. *Teaching and Teacher Education*, 24(4), 982–1002.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 12

Regulation Activities of Teachers in Secondary Schools: Development of a Theoretical Framework and Exploratory Analyses in Four Secondary Schools Based on Time Sampling Data



Katharina Maag Merki, Urs Grob, Beat Rechsteiner, Andrea Wullschleger,
Nathanael Schori, and Ariane Rickenbacher

12.1 Introduction

Previous research revealed that teachers' and school leaders' regulation activities in schools are most relevant for sustainable school improvement (Camburn, 2010; Camburn & Won Han, 2017; Hopkins, Stringfield, Harris, Stoll, & Mackay, 2014; Kyndt, Gijbels, Grosemans, & Donche, 2016; Messmann & Mulder, 2018; Muijs, Harris, Chapman, Stoll, & Russ, 2004; Stringfield, Reynolds, & Schaffer, 2008; Widmann, Mulder, & Köning, 2018). Regulation activities are (self-)reflective activities of teachers, subgroups of teachers, or school leaders that are aimed at improving current practices and processes in classes and in the school in order to achieve higher teaching quality and more effective student learning. Schools that are highly effective in improving teaching and student learning are those that are able to implement tools and processes on an individual, interpersonal, and school level that enable the school actors to think about and adapt current strategies and objectives, to anticipate new possible demands and develop strategies for meeting the demands successfully in the future, and to reflect upon their own adaptation and learning processes. Regulation activities are interwoven in everyday school practices.

However, there are severe shortcomings of previous research, on both a theoretical and a methodological level. For one, there is a lack of a comprehensive theoretical framework to understand regulation in schools, since current models only focus

K. Maag Merki (✉) · U. Grob · B. Rechsteiner · A. Wullschleger · N. Schori
A. Rickenbacher
University of Zurich, Zurich, Switzerland
e-mail: kmaag@ife.uzh.ch

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_12

257

on limited aspects of the regulation activities of teachers and school leaders, and the complex hierarchical and nested structure of everyday school practices has not been considered sufficiently. For another, apart from a few exceptions (e.g. Spillane & Hunt, 2010), research on school improvement and on teachers' formal and informal learning has mostly used self-report on standardized questionnaires, such as teacher surveys on cooperation or teaching practices. The validity of these self-report ratings is limited, however, if the aim is to gain insights into everyday school practices, which is crucial for studying teachers' regulation in the context of school improvement in terms of its significance for student learning (Ohly, Sonnentag, Niessen, & Zapf, 2010; Reis & Gable, 2000).

Hence, in this paper, we develop a framework for understanding regulation in the context of school improvement. Furthermore, we present the results of a mixed-method case study in four lower secondary schools, in which we analysed teachers' regulation activities by using time sampling data of teachers' performance-related and situation-specific day-to-day activities over 3 weeks.

This new methodological approach extends previous research significantly in four different ways: First, whereas in former research teachers' activities were recorded retrospectively, often after a longer period of time, we investigated activities on each day over 3 weeks. This reduces the danger of errors or biases in teachers' remembering of past activities and allows more valid identification of teachers' regulation activities (Ohly et al., 2010; Reis & Gable, 2000). Second, in contrast to investigating activities on a more general level by using self-reports, e.g. at the end of the year, this approach allows us to capture topic-specific activities each day, including informal and formal settings, since a detailed catalogue of activities was provided that helped the teachers to differentiate between the single activities during the day. Furthermore, the approach allows identification of day-specific variation in regulation activities. Third, since the teachers had to specify whether they conducted the activities alone or together with others, the approach allows analysis of the social structure of the regulation activities in a more detailed manner. And finally, since the regulation activities were analysed every day, the relation between day-to-day variation in regulation activities and day-to-day variation in the benefits of these activities for school improvement can be analysed.

In the paper, we first discuss the theoretical background and provide a definition of regulation in the context of school improvement. Second, we present the research questions and hypotheses, followed by a description of the study and the research design. Finally, initial results are presented. The paper closes with a discussion of the strengths and limitations of this newly implemented approach and suggestions for further research.

12.2 Theoretical Framework on Regulation in the Context of School Improvement

12.2.1 Regulation in the Context of School Improvement: Theoretical Anchors

From a theoretical perspective, different approaches exist for describing regulation pertaining to school development. First, of particular interest are approaches that consider the hierarchical as well as the nested and loosely coupled structure of school organisations (Fend, 2006; Weick, 1976) and, in doing so, differentiate between individual and collective regulation processes and activities. Second, due to the dynamic perspective of school improvement (Creemers & Kyriakides, 2012), theoretical approaches have to be able to focus on the processes of regulation.

Accordingly, the present study refers to Argyris and Schön's (1996) theory of organisational learning as a basic theory for understanding individual and collective learning in organizations. As this theory is unspecific in terms of type of organisation, Mitchell and Sackney's (2009, 2011) theory of the learning community is also important for an understanding of individual and collective learning processes particularly in schools. However, neither of the two theories are really able to describe the respective learning processes and learning activities very well. Therefore, self-regulation theories (Hadwin, Järvelä, & Miller, 2011; Panadero, 2017) and particularly the theory of self-regulated learning by Winne and Hadwin (2010) are relevant for this study. The following table (see Table 12.1) provides a brief overview of the core assumptions and theoretical approaches that will be presented subsequently in more detail.

With reference to the first criterion, the theory of organisational learning by Argyris and Schön (1996) and the theory of the learning community by Mitchell and Sackney (2009, 2011) have been crucial for the present study. These theories

Table 12.1 Theoretical anchors for the analysis of regulation in the context of school improvement

Theoretical approach	
Theory of organisational learning (Argyris & Schön, 1996)	Individual and collective learning, including <ul style="list-style-type: none"> • Single-loop learning: change of actions and strategies • Double-loop learning: change of school-related objectives, strategies, and assumptions • Deutero-learning, meta-learning: change of the learning system
Socio-constructivist learning theories, theory of the learning community (Mitchell & Sackney, 2009, 2011)	Individual, interpersonal, and organisational strategies of reconstruction, deconstruction, and construction of knowledge
Self-regulation theories (Hadwin et al., 2011; Panadero, 2017; Winne & Hadwin, 1998)	Regulation strategies of <ul style="list-style-type: none"> • Conditions (tasks and cognition) • Operations • Standards

assume that changes in organizations cannot be explained through individual learning processes of particular actors alone: To a significant extent, changes also involve collective or organisational learning. In contrast to Argyris and Schön's (1996) theory, which can be understood as a basic theory of organisational learning, Mitchell and Sackney's (2009, 2011) theory of learning communities is based on schools explicitly. It, therefore, puts a stronger focus on pedagogical processes and people's growth and development than theories of learning organisations do (Mitchell & Sackney, 2011, p. 8). This is of particular relevance for the study at hand, which we conducted at secondary schools. Mitchell and Sackney (2011) differentiate collective learning processes even further and distinguish between interpersonal and organizational learning processes. This differentiation is crucial for the understanding of schools, since schools are distinguished by their complex structure, ranging from individual teachers to different formal and informal social subgroups and sub-processes that are only loosely coupled (Weick, 1976) to the school's organization as a whole. To understand teachers' regulation in secondary schools, it is necessary to combine these subsystems explicitly so as to increase the ecological validity of the theory.

Accordingly, in this study, we will differentiate between individual regulation (for example, analysis and adaptation of individual lessons by a teacher), interpersonal regulation (for example, analysis of teamwork by a subgroup of teachers and adaptation of the modus of working), and organisational regulation (for example, adaptation of teaching processes based on the results of external evaluation by the school as a whole).

However, the analysis of regulation, regardless of whether the regulation is done by individuals, subgroups of teachers, or the whole school, requires a dynamic perspective on the research topic. This means referencing theoretical concepts that are able to identify and describe the corresponding processes.

As with the first criterion, for understanding regulation as a process, a first important theory is Argyris and Schön's theory of organisational learning (1977, 1996). At the centre are the theories-in-use of the various actors and of the organization. The theory-in-use is the actors' implicit knowledge about the organization, which affects the actors' subsequent actions and their individual and organizational learning. Individual and organizational learning processes are based on a cybernetic model. In the model, actions, objectives, and the learning system as a whole are analysed in a regulatory circle, distinguishing between three different learning modes: (a) single-loop learning, or "instrumental learning that changes strategies of action or assumptions underlying strategies in ways that leave the values of a theory of action unchanged" (Argyris & Schön, 1996, p. 20), (b) double-loop learning, or learning that "results in a change in the values of theory-in-use, as well as in its strategies and assumptions" (p. 21), and (c) deutero-learning (also called second-order learning, or learning how to learn) that enables the members of an organization to "discover and modify the learning system that conditions prevailing patterns of organizational inquiry" (p. 29). The driving force behind these learning processes are challenges or unsatisfactory results, based on which alternative actions and objectives are extrapolated, and the organizational theory-in-use is modified.

For the present study, this means that regulation in schools could be understood as strategies of analysing and adapting current actions in the classroom by individuals, by subgroups of teachers, or by the whole school by reacting to internal or external challenges, conditions, and requirements (single-loop learning). In addition, regulation in schools can be understood as individual and collective strategies of analysing and adapting objectives and values in the school as well as the school's tactics and assumptions (double-loop learning). And finally, regulation is related to analyses of the organization's learning system and the effectiveness of the implemented single-loop and double-loop learning strategies, respectively (deutero-learning).

Although Argyris and Schön's theory is relatively old and learning processes are described in little detail, there are some congruences with current self-regulation theories (e.g. Panadero, 2017; Winne & Hadwin, 2010; Zimmerman & Schunk, 2001, 2007). As do Argyris and Schön, they refer to theories on information processing as well as socio-constructivist learning approaches (Panadero, 2017; Zimmerman, 2001). However, self-regulation theories describe regulation explicitly and in a more differentiated manner (Panadero, 2017; Zimmerman, 2001). These theories assume that learning is a result of an active and (self-)reflective manner concerning information processing; cognitive, metacognitive, motivational-emotional, and resource-oriented learning strategies are applied when dealing with the individual characteristics of the students and the characteristics of the task to be carried out. Further, there is a strong focus on the aspect that knowledge is constructed and thus constitutes a mental representation, which is analysed and advanced through active involvement of the student or teacher depending on the sociocultural and situative context (Järvenoja, Järvelä, & Malmberg, 2015).

The recursive model of self-regulated learning by Winne and Hadwin (1998), which strongly emphasizes (meta)cognitive processes, is of particular relevance for the present study. At its core are five dimensions, abbreviated as COPES: conditions, operations, products, evaluations, and standards. Regulation refers to the three dimensions conditions, operations, and standards. That means that based on an evaluation of the achieved products, either the conditions, the operations, or the standards will be regulated if the achieved products do not fulfil the requirements.

- First, regulation can refer to the conditions of the learning process; these are characteristics of the tasks to be processed (e.g. task resources, time, social context) as well as individual requirements (e.g. beliefs, dispositions, motivational factors, domain knowledge, knowledge of tasks and of study tactics and strategies). In the school context, these comprise, for example, analysis and adjustment of the time available (e.g. provide extra time) to conduct school improvement projects (task conditions) or regulation of teachers' and school leaders' knowledge of school development processes and school improvement strategies by collecting more information on how to proceed effectively (cognitive conditions).
- Second, regulation can refer to the operations that are used for analyzing and processing the available information. Here, cognitive, metacognitive,

motivational-emotional, and resource-oriented regulation strategies can be differentiated. Cognitive strategies in the school context may be, for example, strategies of teachers, a subgroup of teachers, or a steering committee, for summarizing and structuring different school-related pieces of information gained from internal and external evaluations. Metacognitive strategies are, for example, strategies of a subgroup of teachers for analyzing strengths and weaknesses of a new teaching model and for mapping out its further development (Pintrich, 2002). Motivational-emotional regulation strategies are used to increase the teachers' interest in implementing school-related reforms (Järvelä & Järvenoja, 2011; Wolters, 2003). Therefore, school-specific regulation referring to operations can be seen if teachers or groups of teachers analyze and adjust their cognitive, metacognitive, or motivational-emotional and resource-oriented regulation strategies in order to achieve a better understanding of the problem or to increase teachers' motivation to deal with daily challenges.

- Third, regulation refers to the standards that should be achieved. In the school context, corresponding regulation processes are visible if individual teachers, subgroups of teachers, or the entire school modify the standards of a school reform due to difficulties, by, for example, lowering the standards or setting different priorities.

Apart from the approaches by Argyris and Schön (1996) and by Winne and Hadwin (1998), Mitchell and Sackney's (2009, 2011) theory of the learning community is especially interesting for the relevant issues in this study because it provides a pedagogical and multilevel perspective on learning and regulation processes in schools. The theory is again based strongly on a socio-constructivist theory on individual and collective learning. However, it does not emphasize information processing approaches of learning. Mitchell and Sackney (2011) interpret knowledge and knowledge construction as "a natural, organic, evolving process that develops over time as people receive and reflect on ideas in relation to their work in the organization" (p. 40). Based on this approach, school-related regulation can be described as an individual but also collective strategy of active and reflective constructing of knowledge, whereas professional narratives of individuals and groups are reconstructed and deconstructed in a complex process. In doing so, teachers not only deal with their own ideas and experiences and identify their existing practices, reflect on strengths and weaknesses in their work, and "search for one's theory of practice" (p. 21), but also look for new ideas and new knowledge: They discuss new approaches or strategies with others or experiment with new methods and actively seek out new ideas within and outside their school, in order to utilize them for further developing lessons and learning. In the course of this, the objective is the "transition from familiar terrain to new territory" (p. 47).

Mitchell and Sackney's theory (2009, 2011), which also explicitly includes collective regulation strategies, is of particular relevance for this study, since sense-making processes of the actors in organisations have a pivotal effect on their actions (Coburn, 2001; Weick, 1995, 2001). But the theory also highlights social contexts and social interactions in particular as being a key area of influence regarding

learning processes. As a consequence, learning takes place in social interactions, and knowledge – such as knowledge on effective teaching or school development – is reconstructed and deconstructed and thus extended on the basis of previous experiences and knowledge through sensemaking and (meta)cognitive adaptation processes.

12.2.2 Definition of Regulation in the Context of School Improvement

Considering the theoretical references outlined in the previous section, regulation in the context of school improvement can be defined as the (self-)reflective individual, interpersonal, and organizational identification, analysis, and adaptation of tasks, dispositions, operations, and standards and goals by applying cognitive, metacognitive, motivational-emotional, and resource-related strategies. Regulation means to reconstruct and deconstruct the current practices and, subsequently, to further develop the practices by searching for and constructing new knowledge in order to increase the support and learning success of the students. Regulation is a complex, iterative, non-linear, exploratory, and socio-constructive process of dealing with tasks, of which the actions, motivations, emotions and cognitions are recursively related to each other. Regulation can be realised in formal and informal settings (Kyndt et al., 2016; Meredith et al., 2017; Vangrieken, Meredith, Packer, & Kyndt, 2017) and individually or in smaller or larger groups (Hadwin et al., 2011) together with people and institutions from within the school or from outside. Therefore, regulation can be understood as a socially constructed and shared but also socially situated process, since regulation always takes place in social learning situations (Järvelä, Volet, & Järvenoja, 2010; Järvenoja et al., 2015) and is embedded in daily routines (Camburn, 2010; Camburn & Won Han, 2017; Day, 1999; Day & Sachs, 2004; Gutierrez, 2015).

Four different regulation areas can be distinguished: (a) tasks, (b) goals and standards of tasks, (c) dispositions of actors or group of actors, and (d) operations (see Fig. 12.1):

- (a) Tasks are understood in their broad sense. They encompass requirements and challenges for teachers, subgroups of teachers, school leaders, and other actors that arise in the development of the school and teaching and in the support of students. There are, for example, organizational and administrative tasks, tasks in curriculum development, tasks in the development of teamwork, or school-related quality management and development tasks. Consequently, tasks may vary regarding their complexity, instructional cues (e.g. well- vs. poorly-defined tasks), time needed, resources available or regarding who is in charge of carrying out the task (individuals, subgroup of teachers, school leader, or the whole school). Regulation of tasks means to analyse the task that has to be carried out, to make sense of the task or to identify challenging or easier aspects of

Context: (social) structures (individual, cooperative), situations, guidelines	
<p>Tasks of individuals, subgroups, and the school as organisation</p> <p>Content</p> <p>Resources (e.g. time, material)</p> <p>Instructional cues</p> <p>Complexity</p> <p>Social context</p>	<p>Goals and standards of individuals, subgroups, and the school as organisation</p> <p>Complexity</p> <p>Relevance for student learning</p> <p>Differentiation</p> <p>Leeway</p> <p>Alignment</p>
<div style="border: 1px solid black; border-radius: 15px; padding: 5px; display: inline-block;"> <p>Focus of Regulation</p> </div>	
<p>Dispositions of individuals, subgroups, and the school as organisation</p> <p><i>Motivational-emotional:</i> motivational orientations, emotions, mindset, values</p> <p><i>Cognitive:</i> tactics and strategies, declarative and procedural knowledge of school improvement, of task, of strategies, of actor(s)</p>	<p>Operations of individuals, subgroups, and the school as organisation</p> <p>Explicitness vs. implicitness</p> <p>Fit</p> <p>Grain size</p> <p>Depth of the analysis</p>
Individual and collective characteristics: gender, professional experiences, etc.	

Fig. 12.1 Focus of regulation in the context of school improvement

realization of the task, to search for new knowledge to understand the task, and to extend or reduce the complexity of the task, for instance, if the task is too hard to be resolved.

- (b) Goals and standards of tasks in the context of school improvement are closely related to the task that has to be performed. The goals can differ in their complexity (e.g. rather low [organize a meeting] vs. rather high [the introduction of sitting in on classes or new teaching methods]) and in the relevance for supporting students' learning. Further, they can differ in the level of differentiation (e.g. how precisely the goals are described), in their alignment with guidelines, in the leeway to modify the goals, and in the standards that have to be achieved. Regulation of standards and goals means to analyse the appropriateness of the goals of a specific task and the standards that are related to the realization of the task. If necessary, goals have to be modified, extended or diminished, and standards can be lowered or raised to increase the chance of successfully achieving the goals.
- (c) Dispositions of actors or group of actors are relevant conditions for task realization. Motivational-emotional and (meta)cognitive dispositions can be distinguished. The regulation of these dispositions means, for instance, that strategies are applied to increase motivation to deal with the task (e.g. individual and

collective self-efficacy, mindset), to reduce fear or pressure to perform, or to increase knowledge of the task or of the required tactics and instruments to resolve the task.

- (d) Operations are implicit and explicit tactics, methods, and strategies that refer to two different areas: (i) strategies to carry out tasks in schools (e.g. teaching methods, strategies to support students, strategies to cooperate), and (ii) strategies to regulate current practices in schools (e.g. cognitive or metacognitive strategies). In the former, operations may be regulated by making the applied methods and strategies more explicit or by analysing how well the strategies fit for accomplishing the goals of the operations. In the latter, understanding operations as strategies to regulate practices in school, the regulation of these operations means to regulate the analyses, and adaption process itself, or, in the sense of Argyris and Schön (1996), the individual or collective learning system (deutero-learning). Therefore, actors may modify and adjust the ‘grain size’ of the applied regulation strategy, realizing, for instance, that they have been applying overly narrow strategies to deal with teaching problems and that they need to take a wider look at the problem, for instance, by seeking to gain knowledge from experienced teachers outside the school. Further, they might modify the applied regulation strategies by increasing the depth of their analyses to better understand the task.

This understanding of regulation is compatible with the concept of reflective practice or reflection as it is used in many previous studies (Nguyen, Fernandez, Karsenti, & Charlin, 2014; Schön, 1984). As analysed in the systematic review by Nguyen et al. (2014) on theoretical concepts on reflection, regulation is an explicit process of becoming aware and making sense of one’s thoughts and actions with the view to changing and improving them. It is also compatible with the concept of reflective dialogue, which has been identified as a central feature in professional learning communities (Lomos, Hofman, & Bosker, 2011; Louis, Kruse, & Marks, 1996). We also see some congruence between our concept of regulation and the concept of informal learning or workplace learning (Kyndt et al., 2016). These theoretical approaches are interesting for the present model, since they put a focus on everyday learning that occurs not only in formal settings like vocational training but also in not planned and formally structured occasions that are embedded in daily work. However, the concept of regulation developed here represents a significant extension: It is more differentiated than the concepts mentioned, since it explicitly emphasizes the particular regulation practices that help people to understand and to improve current practices. Further, it introduces a multilevel perspective that takes into account the complex, hierarchical, and nested structures of schools as organizations. With this, it will become possible to develop a deeper understanding of regulation in the context of school improvement, to identify possible difficulties in dealing with complex school-related requirements, and to develop approaches for promoting regulation in schools more effectively.

In this paper, an emphasis is put on the analysis of the regulation tasks that are performed over 3 weeks. Of special interest are what daily regulation activities of

the teachers occur, and to what extent possible variabilities are associated with teachers' daily experienced benefit, teachers' daily satisfaction, and teachers' individual characteristics.

12.3 Previous Research on Daily Regulation in Schools and Research Deficits

Research on teachers' regulation in schools has focused above all on the analysis of teachers' reflective practices and on informal learning in the workplace. Studies on teachers' reflective and informal practices have been conducted primarily in three areas: (a) frequency level, or content of the reflection and informal learning on the basis of standardized surveys, qualitative data, or a mixed-method design; (b) efficiency of targeted interventions or professional learning programmes on teacher's reflection and informal learning and identification of significant prerequisites for reflective and informal learning; and (c) efficiency of teachers' reflective practices and informal learning regarding the professionalisation of the teachers, teaching development, or student performance. The studies frequently pursue multiple objectives, although there is a stronger focus on the first two aspects, and research is very much limited in terms of the analyses of effects of reflective and informal practices (Kyndt et al., 2016).

Camburn and Won Han (2017) reanalysed three large US studies comparatively. Taken together, approximately 400 schools with 7500 teachers were analysed using standardized surveys on reflective practices. The results, which were based on teachers' retrospective assessment of their practices, showed that the majority of teachers reported active reflective practices in various forms. However, if the specific contents of reflection are focused on teaching or school-related aspects, for example, the results showed that only some teachers, generally less than half, engaged more frequently in reflective practices. In particular, reflective practices were reported regarding content or performance standards, reading/language arts or English teaching, teaching methods, curriculum materials or frameworks, and school improvement planning. In contrast, reflective practices that would require a considerable amount of introspection and initiative were rather rare (p. 538) (see also Kwakman, 2003).

There were major differences to be found in teachers' reflective practices (Camburn & Won Han, 2017). The differences could be explained particularly by the teachers' experience in reflection or by provision of instructions for professional development. Individual characteristics such as gender or ethnic background seemed to have no effect on teachers' reflective practices. However, the role that the teachers take in schools (e.g. senior managers, teachers, support staff) and the subject that the teachers teach were revealed to be significantly related to teachers' profile of learning (Pedder, 2007).

Besides teachers' individual factors, particularly interest and motivation for reflexive learning, school factors are most relevant for explaining differences between teachers in their reflexive practices, particularly teachers' autonomy, embedded learning opportunities, school culture, support, or leadership (Camburn, 2010; Kyndt et al., 2016; Oude Groote Beverborg, Slegers, Endedijk, & van Veen, 2017).

As to school differences, Camburn and Won Han (2017, p. 542) found hardly any differences in the frequency of reflective practices. The largest difference between the schools was whether or not reflective practices were implemented with the help of experts from outside the schools. However, Pedder (2007) suggested that there are differences between schools if the mix of learning profiles of teachers (e.g. high levels of in-class and out-of-class learning vs. low levels of in-class and out-of-class learning) are identified, analysed by using cluster analyses considering four types of learning (enquiry, building social capital, critical and responsive learning, and valuing learning).

Gutierrez (2015) analysed the reflective practices of teachers as well but, in contrast to Camburn and Won Han (2017), over an entire school year on the basis of a qualitative design. Further, the study aimed to record not only the frequency of reflection over the school year but also the level of reflection. The focus was on the reflective practices of three groups of public school elementary science teachers taking part in a professional development programme. The researcher used a variety of methods, including daily reflective logs, field notes, survey forms, and audio- and video-taped recordings of all the teachers' interactions, which at the same time recorded teachers' reflections on their practice. Through the analysis of reflective interactions, Gutierrez was able to identify three levels of reflective practice: descriptive, analytical, and critical reflection. The levels differed in their complexity (consideration of possible arguments for understanding of situations). Critical reflection was identified as the highest level. Reflective interactions were observed in practically all conversations, but the level of reflection varied in frequency. Descriptive reflective interactions were the most frequent (43%), followed by analytical (30.8%) and critical reflective interactions (26.2%). Further, reflective practice was less visible in normal conversations but was especially visible where it was initiated by "knowledgeable others" (Gutierrez, 2015).

A look at Gutierrez (2015) and Camburn and Won Han (2017) yields the insight that less complex reflective practices take place more often than more complex reflective practices. This is also evident in the German-speaking context (Fussangel, Rürup, & Gräsel, 2010; Gräsel, Fussangel, & Parchmann, 2006; Gräsel, Fußangel, & Pröbstel, 2006), which is also the context in which the study presented here was conducted. However, the two studies also found that reflective practices can be facilitated by selected external persons, "knowledgeable others" (Gutierrez, 2015) or "instructional experts" (Camburn & Won Han, 2017), which is in line with various other studies on the professionalisation of teachers and school development (Butler, Novak Lauscher, Jarvis-Selinger, & Beckingham, 2004; Creemers & Kyriakides, 2012; Day, 1999; Desimone, 2009; Kreis & Staub, 2009, 2011; West & Staub, 2003).

Even though these studies provide some insights on teachers' reflective practices and informal learning, various questions remain open concerning both methodology and content. Whereas the methodological approach chosen by Gutierrez (2015) or others (see e.g. Raes, Boon, Kyndt, & Dochy, 2017) allows for a simultaneous recording of reflective activities without the bias of individual distortion through retrospective recording, the approach can only be used in small samples because of the time requirements for data collection. In contrast, it is possible to gain insights into the reflective activities of a large number of teachers using the standardized approach chosen by Camburn and Won Han (2017); however, these insights are restricted in their validity because of self-reports, since they constitute reflective actions that are evaluated retrospectively and interpreted subjectively. This presents similar methodological difficulties to those that have been discussed in self-regulation studies for years (e.g. Spörer & Brunstein, 2006; Winne, 2010; Wirth & Leutner, 2008).

Since research on teachers' reflection and informal learning is basically dominated by qualitative approaches that allow exploratory gathering of in-depth knowledge on professional learning but are limited in terms of generalisation of the results (Kyndt et al., 2016), new approaches with a more quantitative perspective have to be developed. These approaches should be effective in assessing how teachers regulate their work in a daily situation concretely, taking into account a more dynamic perspective and how effective the regulation strategies are for teachers' and students' learning (see also Oude Groote Beverborg et al., 2017, and the paper in this book). Therefore, analysis of teachers' day-to-day practices and learning requires methods that are able to record individual activities as promptly and accurately as possible. This would not only increase the ecological validity of the measurements but would also aid progress in the development of a theoretical understanding of regulation in the context of school improvement.

In classroom research, strategies with daily logs for teachers have been developed in recent years that make it possible to record concrete day-to-day classroom practices (Elliott, Roach, & Kurz, 2014; Glennie, Charles, & Rice, 2017). Corresponding analyses have revealed that in this way, interesting insights into concrete classroom practices can be gained – insights that systematically increase the level of knowledge and are associated systematically with external criteria, such as with student performance – and that these methods can be deemed valid based on comparison with other methods, such as observations (Adams et al., 2017; Kurz, Elliott, Kettler, & Yel, 2014).

In school development research as well, there are initial studies available that assessed performance-related activities and practices using various methods. Accordingly, studies by Spillane and colleagues analysed the daily activities of school leadership based on experience sampling data (Spillane & Hunt, 2010) and end-of-day log data (Camburn, Spillane, & Sebastian, 2010; Sebastian, Camburn, & Spillane, 2017). In addition, interviews, observation data, or standardized surveys were used. The studies found a high variability in the activities of the school leaders (e.g. administration, instruction) and also substantial differences between the respective school leaders as well as in the course of the week. According to Spillane

and Hunt (2010), three types of school leaders' practices can be differentiated: administration-oriented, solo practitioners, and people-centred.

Sebastian et al. (2017) found that the variation in school leadership practices is domain-dependent, whereas differences were particularly significant for the domains "student affairs" or "instructional leadership" and particularly small for the domains "finances" or "professional growth." In the course of a week, there were only a few differences. One of these differences concerned individual development ("professional growth"): These activities seemed to be performed at the end of the week rather more often, whereas other tasks (e.g. community/parent relations and instructional leadership) were less likely to be performed at the end of the week. The differences between school leaders could be attributed to a (weak) influence of the school's performance level as well as size and type of school.

Further, the analyses showed that valid information on school improvement processes can be gained regarding the daily activities of school leaders with the help of the chosen methods (Camburn et al., 2010; Spillane & Zuberi, 2009). Moreover, a comparison between experience sampling methods and daily log methods showed that both methods delivered similar results; however, the daily log method has proven to be easier in its application and less intrusive on a daily basis (Camburn et al., 2010).

Johnson (2013) investigated school development activities as well. The study analysed 18,919 log entries of instructional coaches at 26 schools, who supported the schools in meeting the needs of at-risk and low-income students (the sample included 23 Title I and three School Improvement Grant schools in the Cincinnati Public Schools). Their specific activities were subsumed under three different categories, and the study analysed to what extent the patterns of categories of work were connected to different state performance indicators. In addition, the results showed that differences in the activities of the school leaders can be identified based on the chosen methods, which, furthermore, correlated with performance indicators.

In summary, research has found that more differentiated information on the activities of teachers, school leaders, and coaches can be gathered using the daily log method rather than with retrospective methods. In contrast to the studies referred to above, what is still missing in the literature are studies that assess the teachers' daily regulation activities outside the classroom with the help of daily logs. Therefore, it remains unclear to what extent teachers deal with their concrete work reflectively and to what extent they regulate it.

Hence, the goal of the case study presented here is to describe the regulation activities of teachers at four secondary schools over 3 weeks. With reference to the theoretical framework presented in Sect. 12.2.2, the main focus is on the regulation of tasks, e.g. organisational-administrative tasks, teaching and learning tasks, or team and school development tasks. However, we will not be able to analyse what regulation strategies the teachers applied, or on what quality level they regulated these aspects. Therefore, we will not corroborate the validity of the theoretical framework. Instead, our first aim is to obtain insights into the day-to-day regulation activities of teachers at secondary schools and to extend the respective literature particularly by analysing teachers' day-to-day activities. To achieve this, we

developed a new task- and day-sensitive instrument for teachers that is based on the time sampling method (Ohly et al., 2010; Reis & Gable, 2000). Our second aim is to investigate the validity of the instrument. However, one has to keep in mind that only a small school sample is examined. Therefore, the analyses can be interpreted only as exploratory.

12.4 Research Questions and Hypotheses

To achieve the aims of this study, we analyse two different sets of research questions: The first set of questions examines teachers' daily regulation activities and analyses differences between tasks, parts of the week, persons, and schools. To investigate the validity of the newly developed instrument, we test hypotheses related to previous research. The second set of questions examines the relation between teachers' daily regulation activities and teachers' perceptions of the benefit of these activities for student learning, teaching, teacher competencies, and team and school development. Further, we investigate the associations with teachers' daily satisfaction. Again, to verify the validity of the instrument, we test hypotheses based on previous research.

Set of Questions No. 1: Daily Regulation Activities

Question 1a: What daily regulation activities occur in the participating schools, and what is their frequency?

Hypothesis 1 (H1): In particular, the greater part of regulation activities is expected to relate to teaching classes and to administrative-organisational tasks. Regulation activities that require a higher level of introspection and initiative are conducted considerably less frequently, however (Camburn & Won Han, 2017).

Question 1b: To what extent do the daily regulation activities during the week (from Monday to Friday) differ from daily regulation activities on the weekend?

Hypothesis 2 (H2): Systematic differences are expected (Sebastian et al., 2017): Activities that require on-site interactions (e.g. teaching, meetings) will take place during the week more often than on weekends. Moreover, regulation activities that are closely related to demanding situations at school and that require activities in a timely and – if necessary – collaborative manner with other teachers are expected to occur more often on weekdays than on the weekend. Class preparation or follow-up activities are expected to take place on a similar relative level on weekdays and on weekends, since teachers often do teaching preparation or grade student work also on the weekend. In contrast, teachers' study of specialist literature is expected to occur relatively more often on weekend days than during the week as teachers have more free time on weekend days.

Question 1c: To what extent are there differences among the schools in selected regulation activities specifically relevant for school development?

Hypothesis 3 (H3): We expect to find differences among schools (Camburn & Won Han, 2017; Pedder, 2007; Sebastian et al., 2017). However, since only four schools participated in this case study, we expect to find only small differences.

Question 1d: To what extent are there differences among teachers?

Hypothesis 4 (H4): Systematic differences among teachers have to be assumed (Camburn et al., 2010; Camburn & Won Han, 2017; Pedder, 2007; Sebastian et al., 2017; Spillane & Hunt, 2010).

Hypothesis 5 (H5): Teachers with specific leadership roles in schools (e.g. school leader, member of a steering committee) differ from teachers with no leadership roles in particular areas (Pedder, 2007; Sebastian et al., 2017). For example, it can be expected that teachers with leadership roles are involved in activities concerning school quality and school development more often than teachers with no specific leadership roles, and that they are more likely to carry out tasks on behalf of the school. Regarding class teachers with a special responsibility for their classes, a special focus concerning reflection upon their own teaching practices is expected.

Set of Questions No. 2: Interrelation Between Daily Regulation Activities, Perceived Benefit, and Level of Satisfaction

Question 2a: How do teachers perceive the benefits of the daily regulation activities, and how satisfied are teachers at the end of the day? To what extent are there differences among the schools?

Hypothesis 6 (H6): In Switzerland, the main focus of teacher training and continuing education is on improving competencies in the area of teaching and learning. In contrast, competencies in the area of team and school development are promoted less purposefully (Schweizerische Konferenz der kantonalen Erziehungsdirektoren, 1999). Therefore, it is expected that teachers are able to realize their daily activities in a particularly beneficial manner regarding student learning but to a lesser degree when it comes to team and school development. With this in mind, it can also be assumed that teachers' perceived benefit of the activities will be higher for supporting student learning than for team and school development.

Hypothesis 7 (H7): Systematic differences are expected between the schools, since schools differ significantly in their school improvement capacity (e.g. Hallinger, Heck, & Murphy, 2014). As with hypothesis 3 (H3), however, we assume only small differences between schools.

Question 2b: To what extent are teachers' daily regulation activities related to teachers' daily perceptions of benefit and teachers' daily satisfaction level?

Hypothesis 8 (H8): Teachers' regulation activities realized during the day are related systematically to the perceived benefit (H8a) and the level of satisfaction at the end of the day (H8b). However, the strength of the associations between regulation activities and perceived benefit can vary depending on the strength of the overlap between the content of the regulation activities (e.g. individual teachers' reflection upon and further development of their teaching) and the area of bene-

fits (e.g. regarding the improvement of individual teaching practices). As to the relation with the level of satisfaction, previous research is missing. However, we argue in analogy to school improvement and self-regulated research: For instance, school improvement research shows that it is not school leaders themselves but a specific type of leadership that is most beneficial to school improvement (e.g. Hallinger & Heck, 2010). Additionally, the literature on self-regulated learning demonstrates that it is not the quantity itself but the quality of the implemented strategy that is beneficial for learning (e.g. Wirth & Leutner, 2008). Similarly, a rather weak connection between teachers' regulation activities (quantity) and their level of satisfaction at the end of the day is expected.

Question 2c: To what extent is teachers' perceived daily benefit related to their daily level of satisfaction? To what extent do the relations between daily benefit and satisfaction differ among the schools?

Hypothesis 9 (H9): Following the argumentation in hypothesis 8 (H8) above, it is expected that teachers' perceived daily benefit relates systematically to their daily level of satisfaction. This correlation becomes apparent especially when teachers have experienced their daily activities as beneficial for the "core business" of teachers – student learning – and their individual development of teaching practices and competencies (Landert, 2014).

Hypothesis 10 (H10): Correlation strengths between perceived benefit and the level of satisfaction in schools provides information on how important the benefits in a specific area are for satisfaction in a school. Since schools seem to put a focus on teaching and learning processes to different degrees, and since they realize school development processes in different manners (e.g. Hallinger et al., 2014; Muijs et al., 2004), we expect to find differences among the schools in terms of the correlation between teachers' perceived daily benefit and teachers' daily level of satisfaction.

Question 2d: To what extent do individual factors influence the relation between teachers' perceived daily benefit and teachers' daily satisfaction level?

Hypothesis 11 (H11): The expectancy-value theory (Eccles & Wigfield, 2002) assumes that the perceived value of a specific goal as well as the expectation of being able to achieve the goal have an influence on a person's motivation to engage in specific activities. Accordingly, it is assumed that teachers, who are especially interested in analysing and developing teaching and learning processes, are able to benefit more from daily activities that are perceived as beneficial when it comes to their individual levels of satisfaction. They will be especially dissatisfied if their perceived daily activities are deemed less beneficial. Accordingly, we expect to find a closer relation between teachers' perceived daily benefit and teachers' daily level of satisfaction for teachers with a higher level of interest than for teachers with a lower level of interest (moderation effect).

Hypothesis 12 (H12): In contrast, given that there are neither theoretical arguments nor any empirical evidence, it is expected that individual characteristics, such as teachers' sex and length of service, do not have any systematic moderating influ-

ence on the relation between teachers' perceived daily benefit of daily regulation activities and teachers' daily satisfaction levels.

12.5 Methods

12.5.1 Context of the Study and Sample

The study depicted was a mixed-method case study in four lower secondary schools (ISCED 2) in four cantons in the German-speaking part of Switzerland. In these cantons, the compulsory school system is structured into two different levels (primary and lower secondary level), and the total period of compulsory education amounts to 11 years (<http://www.edk.ch/dyn/16342.php>; [June 12, 2018]). Generally, compulsory education starts at age 4. The primary level – including 2 years of kindergarten – comprises 8 years and the lower secondary level 3 years. In lower secondary schools in the cantons, where we conducted the study, several teachers educate the same students. Therefore, they need to exchange materials and information about the students. In addition, special education teachers and social work teachers extend the regular teaching staff at the schools. Due to the assignment of a greater autonomy to the schools, the schools are required to regularly assess the strengths and weaknesses of teaching and the school. Therefore, school improvement and the regulation of school processes are mandatory and are supervised by external school inspections. However, in contrast to other countries, this is only a low-stake, supportive monitoring without a lot of social pressure (Altrichter & Kemethofer, 2015); the schools, school leaders, and teachers do not have to fear severe consequences if they fail to meet the expectations.

All schools participated voluntarily in this study. For the selection of the schools, it was important to be able to consider different school contexts, considering both rural and urban schools as well as schools in communities with a high or low socio-economic level.

In total, 96 of the total population of 105 teachers and school leaders participated (response rate: 91.4%). The sample in the time sampling sub-study was a bit smaller, however. Here, we were able to analyse the data of 81 participants. Correspondingly, the response rate of 77.1% was a bit lower but still very high (School1 = 87.5%, School2 = 65.2%, School3 = 76.7%, School4 = 78.6%). Table 12.2 shows the composition of the sample in terms of sex, workload (in grades), role (combination in four main groups), and schools.

Since all but one school leader also had to teach classes, we use the term 'teacher' for all participants. The average length of service of the 81 teachers was 14.6 years ($SD = 9.2$). Moreover, many of the teachers had been working at the school examined for many years ($M = 10.2$, $SD = 8.2$). There was no significant difference among the four schools in teachers' length of service ($F(3,70) = 0.013$, $p = 1.00$) or in length of service at the current school ($F(3,70) = 0.247$, $p = .86$) (no table).

Table 12.2 Sample for the time sampling sub-study

		n	%
Sex	Female	40	54.1
	Male	34	45.9
	No response	7	
Workload (% FTE)	$0\% \leq x < 20\%$	0	0.0
	$20\% \leq x < 40\%$	3	3.7
	$40\% \leq x < 60\%$	13	16.0
	$60\% \leq x < 80\%$	10	12.3
	$80\% \leq x \leq 100\%$	48	59.3
	No response	7	
Role	Special needs teacher ^a , teacher of German as a second language ^a , therapist ^a	2	2.5
	Subject teacher ^a	29	35.8
	Class teacher ^a	28	34.6
	Leadership role (school leader, steering committee)	22	27.2
School	School 1	21	25.9
	School 2	15	18.5
	School 3	23	28.4
	School 4	22	27.2
	Total	81	100.0

Note. There are no data on sex and work load available for 7 of the 81 participating persons. The percentages refer to valid values; *FTE* full time equivalent

^aWith no leadership role

In total, the very high response rate indicates a very solid empirical data base. Most of the persons who did not take part in the study were on maternity leave or were on a sabbatical from teaching and schoolwork. Therefore, only very few teachers missed filling in the daily practice log. Besides the time sampling sub-study and before the time sampling started, the teachers had to fill in a teacher questionnaire that assessed important dimensions of regulation processes, including interest in and motivation for regulation processes, cognitive and metacognitive regulation strategies, and the school's social and cognitive climate. Further, a network analysis was conducted at each school. However, in this paper, we focus basically on the time sampling data.

12.5.2 Data Collection and Data Base

12.5.2.1 Recording of Regulation Activities

The time sampling method was applied to identify topic-specific day-to-day practices in schools. This method allows more valid identification of teachers' activities than the method of only asking teachers at the end of the year to retrospectively

Table 12.3 Time structure of the on-line journal entries

Week 1	Week 2	Week 3	Week 4	Week 5
Survey (7 days)	No survey (7 days)	Survey (7 days)	No survey (7 days)	Survey (7 days)

report the intensity of their activities (Ohly et al., 2010; Reis & Gable, 2000). In addition, capturing activities and associated ratings has an advantage over a more closely meshed recording of a day's activities (e.g. using experience sampling) in that it is less work for the teachers; they only have to record the activities once a day and not all throughout the day, and there is no substantial loss of validity (Anusic, Lucas, & Donnellan, 2016; Camburn et al., 2010).

During three 7-day weeks between fall and Christmas 2017 (a total of 21 days), teachers' activities were assessed using a newly developed tool. Teachers filled in a daily on-line practices log at the end of each workday (including weekend days if work had been done). There was a week's break between each daily log week in order to reduce teachers' burden and workload (see Table 12.3).

One week prior to the first daily log day, all teachers received a personalized e-mail with information on the procedure and how to log in their activities. They had two options for filling in the daily log: (1) via an internet-based programme on their computer, or (2) via an app on their smartphone. Every day, at 5 p.m., they received a text message or an e-mail with the invitation to log in the activities of the day. They had time until 2 p.m. the next day to do so. Based on numerous reports from teachers that the time window was too small, we extended it by an additional day in the second week of the survey. There were no problems regarding the activities' assignment to a specific day.

Right at the end of the data recording period, we conducted interviews with selected teachers and the school leaders at each school. The interviews revealed that the teachers found it easy to fill in their daily activities log. At the beginning, the daily logging was somewhat unfamiliar, but, after a short time, as the teachers became acquainted with the categories and single steps, they carried out the procedure without any major problems. Further, the teachers confirmed the validity of the newly developed measurement instrument, particularly also the categories provided.

The daily practice log had two parts. In the first part, the teachers had to answer three questions¹:

1. "You are involved in different activities in your school life. Please state for each activity what category you ascribe it to (e.g. teaching)." The teachers had to identify each activity based on a catalogue of four main categories and 15 sub-categories (see Table 12.4). These categories are in line with the official guidelines for school work in Switzerland. To gain an overview on the daily range of activities, any activities that could not be interpreted as primarily regulation activities were also included – especially teaching lessons, class prepara-

¹Only the first question will be analysed in this paper.

Table 12.4 Main categories and sub-categories to identify daily activities (regulation activities shown in bold)

Teaching, support of students and parents
Teaching lessons, incl. break supervision and excursions, special events with class or learning group
Class preparation and follow-up activities, grading, assessing the competencies of the students
Reflecting upon and further developing individual lessons
Talking with students and legal guardians outside of class
Cooperation at team level
Exchange on organisational and administrative questions
Exchange on subject-specific questions
Design and further development of teams/work groups
Collaboration at school level
Participating in quality management and development (e.g. evaluation, school projects, organisation development)
Taking part in school conference meetings
Realisation of tasks for the school (e.g. organising school events, taking over duties)
Professional development
Attending school-internal and -external professional development training
Studying specialist literature
Individual feedback (e.g. sitting in on classes)
Taking part in supervision/intervision

tion and follow-up activities, or talking with students and legal guardians. Regulation activities are highlighted in Table 12.4 in bold type.

2. They (the teachers) had to specify whether they conducted the activities alone or together with others: “Please state for each activity if you performed it alone or together with others.” Possible answers included: alone, with the school leader, with my own team that meets regularly, with special needs teachers.
3. They (the teachers) had to indicate how long the activities lasted: “Please state the approximate duration of each activity.” The response scale was: hours (1 to 8) and minutes (in 10-minute sections: 10 to 50).

In the second part of the daily practice log, the teachers had to rate the benefit of their day in terms of six aspects on a 10-point Likert scale (1 = not at all beneficial, ..., 10 = highly beneficial): “If you think back to the past day as a teacher/expert, how beneficial do you rate this day for the following aspects:

- for reaching the students’ learning goals
- for the best support and promotion of the students
- for the development of my competencies
- for the development of my teaching
- for the development of our work in the team
- for the development of our school.”

Further, they had to rate their day in terms of overall satisfaction and stress,² again based on a 10-point Likert scale (1 = not at all, ..., 10 = extremely): “If you think back on this day as a teacher/expert, how satisfied are you with the day all in all?”, and “If you think back on this day as a teacher/expert, how stressful was this day for you all in all?”

For each teacher, data on up to 21 days were available, resulting in a total of 947 daily records of 81 teachers.

12.5.2.2 Assessment of Interest

For the analysis of possible moderator effects (see research question 2d), two scales were used that were administered through the standardized teacher survey: internal search interest and external search interest.

The scales internal search interest (6 items, Cronbach’s alpha = .78; one-dimensional) and external search interest (6 items, Cronbach’s alpha = .67; two-dimensional) were developed following Mitchell and Sackney’s (2011) concept of internal and external search for knowledge. Internal search interest included to what extent teachers have a substantial interest in learning why certain practices do not work well in their classes, how effective their teaching really is, how good their students really are, and what can be improved in class. An example item for internal search interest was: “Teachers (...) differ according to their interests. To what extent are you (...) interested in different topics? Please state what you (...) would absolutely like to know for your professional daily routine: Absolutely knowing why certain teaching practices do not work well in your own class.”

In contrast, the external search interest scale included substantial interest on the part of teachers in ascertaining methods or strategies with which other teachers are able to promote the students particularly well or what methods are available for giving fair grades. This scale was two-dimensional: The first dimension referred to interest in expert knowledge, and the second dimension referred to interest in the experiences of other teachers. An example item for external search interest was: “Teachers (...) differ according to their interests. To what extent are you (...) interested in different topics? Please state what you (...) would absolutely like to know for your professional daily routine: Absolutely knowing how other teachers teach.” Teachers responded to these statements on a 6-point Likert scale from 1 (strongly disagree) to 6 (strongly agree).

²Only the question about overall satisfaction is analysed in this paper.

12.5.3 *Data Analysis*

To answer the research questions on the frequency of the participating school members' daily activities in the first set of questions, their daily activity data were recorded dichotomously (1 = activity performed this day; 0 = activity not performed this day). Not considered was the extent to which certain activities had taken place more than once a day or the duration of the reported activities. Hence, these transformed activity data bring into light the absolute number of daily occurrences of specific activities as well as their proportion relative to the number of days with any entry of an activity. The data were analysed using multilevel analysis. Day-to-day changes in the activities over the assessed 21 days, respectively the use of time series analysis, were not the focus here.

Differences between activities that took place during the week and activities that took place on weekends (question 1b) were tested statistically using chi-square tests.

Differences among the schools (question 1c) were calculated using binary logistic multilevel analyses based on dummy variables for the schools.

For the analyses on a personal level (question 1d), the information on the daily activities was aggregated person-related across all days. Question 1d was analysed descriptively and, for the analysis of differences between persons with different roles, by means of binary logistic multilevel analyses. Therefore, three groups were compared: (1) class teachers, and (2) subject-specific teachers, both with no leadership roles, and (3) teachers with leadership roles.

The answers to the research questions in the second set on the relation between teachers' regulation activities, perceived daily benefits, and levels of satisfaction were given descriptively on a daily basis (question 2a). Differences among the schools were then examined using linear multilevel analyses (level 1: daily entries, level 2: persons).

The answers regarding research question 2b were given on the level of daily activities using Pearson correlation coefficients between teachers' daily activities and teachers' perceived daily benefits.

To answer question 2c on the relation between teachers' perceived daily benefits for three target areas (students, teachers, team/school) and daily level of satisfaction, correlations were calculated for each school separately, and differences in coefficients were tested statistically using multilevel analyses.

To answer the last question, 2d, on possible influencing factors on a personal level on the relation between teachers' perceived daily benefit and daily satisfaction level, random slope multilevel analyses were used with the slope of each person being explained through their characteristics (here: teachers' interest, their sex, and length of service).

To reduce type I errors, for all but one of the above multiple hypotheses tests, we applied an adjustment of the significance criterion using the Holm-Bonferroni method. The analysis of the last question, 2d, was the exception, since the number of hypotheses was limited, and they should be decided separately upon and not family-wise.

12.6 Results

12.6.1 Set of Questions No. 1

12.6.1.1 What Daily Regulation Activities Occur in the Participating Schools, and What Is Their Frequency? (Question 1a)

The results are compiled in Table 12.5. They show the number of daily entries of different activities and the proportion relative to all days on which any entry was made. The underlying data were structured dichotomously (activity was performed vs. was not performed on a given day).

As expected, activities in teachers' 'core business' areas exhibited the highest relative frequencies. They were: *Class preparation and follow-up activities* (84.1% of entries), *teaching* (71.6%), and somehow less often, *talking with students and legal guardians outside of school*, respectively (27.5%); 40.5% of entries indicated *exchange on organisational and administrative questions*, followed by *reflection on and further development of individual teaching practices* (30.1%), *exchange on subject-specific questions* (23.1%), and *design and further development of teams/work groups* (13.1%). Regulation activities in the area of *school quality management and development* were much rarer (5.4%). Completing tasks for the school was recorded approximately once every seventh day. Finally, one series of activities

Table 12.5 Absolute and relative frequency of different activities (regulation activities shown in bold)

	n	Percentage (%)
Class preparation and follow-up activities	796	84.1
Teaching	678	71.6
Exchange on organisational and administrative questions	384	40.5
Reflection on and further development of individual teaching practices	285	30.1
Talking with students and legal guardians outside of school	260	27.5
Exchange on subject-specific questions	219	23.1
Realisation of tasks for the school	136	14.4
Design and further development of teams/work groups	124	13.1
Study of specialist literature	60	6.3
Further training, both within the school and externally	52	5.5
Participating in quality management and development	51	5.4
Taking part in school conference meetings	43	4.5
Individual feedback (e.g. sitting in on classes)	42	4.4
Taking part in supervision/intervision	9	1.0

Note. Data basis: daily entries (N = 947)

All activity data refer to summed-up occurrences (no: 0/yes: 1) on a day. The percentages represent proportions relative to the total number of days on which at least one school-related activity was reported (N = 947). Multiple responses were possible (column sum of percentages >100%)

exhibited a clearly marginalized status – namely, the hardly occurring *taking part in supervision or intervision* (1.0%), *individual feedback* (4.4%), *taking part in school conference meetings* (4.5%), and *further training both within the school and externally* (5.5%). *Studying specialist literature* was reported approximately every 16th day only.

12.6.1.2 To What Extent Do the Daily Regulation Activities During the Week (from Monday to Friday) Differ from Daily Regulation Activities on the Weekend? (Question 1b)

Out of 947 entries of activities, 813 (85.9%) occurred on a weekday, and 134 activities (14.1%) occurred on the weekend (no table). Hypothetically assuming an equal distribution of activities over all 7 days, five out of seven activities (71.4%) would have been performed during the week and two out of seven activities (28.6%) on the weekend. However, the results revealed that school-related activities on weekends were less frequent than during the week (14% of all activities instead of 28% when assuming equal distribution). Yet, the weekend days were also used for school-related activities, albeit a bit less intensively (Table 12.6).

Table 12.6 Average distribution of different activities on weekdays and on weekends (regulation activities shown in bold)

	During the week (%) ^a	On weekends (%) ^b	p ^c
Class preparation and follow-up activities	85.2	76.9	ns
Teaching	82.8	3.7	p < .001
Exchange on organisational and administrative questions	45.5	10.4	p < .001
Reflection on and further development of individual teaching practices	32.5	15.7	p < .001
Talking with students and legal guardians	31.2	4.5	p < .001
Exchange on subject-specific questions	26.1	5.2	p < .001
Realisation of tasks for the school	15.6	6.7	ns
Design and further development of teams/work groups	14.9	2.2	p < .001
Participating in quality management and development	6.0	1.5	ns
Study of specialist literature	5.8	9.7	ns
Further training, both within the school and externally	5.4	6.0	ns
Taking part in school conference meetings	5.2	0.7	ns
Individual feedback (e.g. sitting in on classes)	4.3	5.2	ns
Taking part in supervision/intervision	1.1	0.0	ns

Note. Sequence organized according to percentage during the week
Multiple responses were possible (column percentage total > 100%)

^aData basis: daily entries for weekdays ($n = 813$)

^bData basis: daily entries for Saturdays or Sundays ($n = 134$)

^cStatistically tested using chi-square tests; significances adjusted using the Holm-Bonferroni method

Table 12.6 documents the *relative* percentages of the 14 activities analysed within all activities on weekdays vs. weekends. It should be noted that an equally high percentage does not signify equally frequent activities on weekdays and on weekends, when viewed absolutely, but rather an *equal percentage* relative to all reported activities on weekdays and relative to all reported activities on weekends.

Teachers used the weekends especially for *class preparation and follow-up activities* (76.9%), followed by *reflection on and further development of individual teaching practices* (15.7%), and by *exchange on organisational and administrative questions* (10.4%), which can be engaged in easily nowadays through electronic means of communication.

Comparing weekdays and weekends, the results revealed logically coherently that the largest differences appeared in activities that were often place- or time-bound, most of all *teaching* (3.7% on weekends vs. 82.8% on weekdays), but also *exchange on organisational and administrative questions* (10.4% on weekends vs. 45.5% on weekdays), *exchange on subject-specific questions* (5.2% on weekends vs. 26.1% on weekdays), or *design and further developments of teams and work groups* (2.2% on weekends vs. 14.9% on weekdays). *Reflection on and further development of individual teaching practices* was also relatively more common on workdays than on weekend days (32.5% on weekdays vs. 15.7% on weekends).

Whereas *further training activities* and *individual feedback* were reported to a similar relative extent on weekends as on weekdays, the *study of specialist literature* had a nominally slightly higher rating on weekends (9.7% vs. 5.8%), which might be attributed to more time being available. However, this difference was not significant (even without Holm-Bonferroni adjustment).

12.6.1.3 To What Extent Are There Differences Among the Schools in Selected Regulation Activities Specifically Relevant for School Development? (Question 1c)

Two forms of activity were chosen for answering the research question on differences between schools in regulation activities. The two activities are of special interest from a school development perspective, and they occur in sufficient frequency: *Reflection on and further development of individual teaching practices* and *exchange on subject-specific questions*. Table 12.7 shows the average activity percentages by school. The binary logistic multilevel analyses with dummy variables for the schools exhibited no significant contrasts, even without Holm-Bonferroni adjustment. The schools did not differ in the relative percentages of the two activities.

12.6.1.4 To What Extent Are There Differences Among Teachers? (Question 1d)

So far, the daily entries for school-related activities constituted the evaluation units ($N = 947$). In the following, we examine how the activities were depicted on a personal level ($N = 81$) and what differences between the teachers could be identified.

Table 12.7 Activities relevant to school development by school

	School 1 n teachers = 21 n entries = 254	School 2 n teachers = 15 n entries = 122	School 3 n teachers = 23 n entries = 229	School 4 n teachers = 22 n entries = 295	p ^a
Reflection on and further development of individual teaching practices	26.0%	27.5%	29.7%	35.0%	ns
Exchange on subject-specific questions	21.8%	29.0%	22.5%	22.2%	ns

Note. ^aStatistically tested using binary logistic multilevel analyses (dummy coding of schools); significance of multiple contrasts adjusted using the Holm-Bonferroni method

Table 12.8 Average distribution of different activities on a personal level (regulation activities shown in bold)

	Average proportion (%)	Standard deviation (%)
Class preparation and follow-up activities	80.1	26.5
Teaching	73.0	21.3
Exchange on organisational and administrative questions	43.7	26.4
Reflection on and further development of individual teaching practices	33.3	30.4
Talking with students and legal guardians outside of school	27.3	25.0
Exchange on subject-specific questions	27.2	24.3
Realisation of tasks for the school	14.0	19.3
Design and further development of teams/work groups	15.6	21.3
Study of specialist literature	5.6	9.1
Further training, both within the school and externally	6.3	14.1
Participating in quality management and development	6.3	13.4
Taking part in school conference meetings	7.7	19.4
Individual feedback (e.g. sitting in on classes)	4.9	16.2
Taking part in supervision/intervision	0.9	5.2

Note. Data basis: percentages of days with specific activity (occurs vs. does not occur) aggregated on a personal level

For this purpose, the daily dichotomous entries for the activities on a personal level were aggregated into average values (see Table 12.8). Person-related, these averages are to be interpreted as frequency percentages of activities on the days documented by each person. For example, if an activity had the value of 33.3%, as was the case with *reflection on and further development of individual teacher practices*, it follows that the 81 teachers on average reported this activity on every third documented day.

The results differed just marginally from the percentages documented in Table 12.5 on the level of daily activities. However, aggregation on a personal level allowed analysis of the differences between persons. Figure 12.2 depicts a series of diagrams that show, with a resolution of 5%, how the activity percentages of the 81 persons were constituted.

The distributions of the average relative frequencies of different activities on a personal level scattered strongly for specific forms of activity. Regarding regulation activities, especially high variances appeared with *exchange on organisational and administrative questions*, *exchange on subject-specific questions*, and *reflection on and further development of individual teaching practices*. Other forms of activity – of course, most of all, activities with a very low absolute response frequency – but also the very widespread *class preparation and follow-up activities*, exhibited far fewer differences or less distribution.

To analyse the relation between daily activities and teachers' school-related roles, we classified teachers into three groups: (1) class teachers, (2) subject-specific teachers, and (3) teachers with leadership roles. Table 12.9 documents the average percentages of the frequency of the 14 different activities by role. As to the regulation activities that are of interest in this context, the results showed that class teachers were involved especially often in the regulation activities *reflection on and further development of individual teaching practices* (together with subject teachers) and *exchange on organisational and administrative questions* (apart from exhibiting a higher percentage of classes taught or *talking with students and legal guardians*). Teachers with leadership roles, however, engaged in *school-related tasks* and *participation in quality management and development* slightly more often.

However, the differences identified resulted from a systematic analysis of all contrasts between the three groups regarding 14 features, i.e. from a total of 42 pairwise comparisons. Because of the multitude of hypothesis tests, the alpha inflation problem arose. When a Holm-Bonferroni adjustment was carried out in order to neutralize this problem, the significance criterion intensified severely. For the contrast with the lowest p -value, the significance threshold would be at $p < .0011$ instead of, uncorrected, .05. With these Holm-Bonferroni adjustments, no contrast exhibited an alpha error below the corrected threshold value. Accordingly, the differences were no longer significant.

12.6.2 Set of Questions No. 2

12.6.2.1 How Do Teachers Perceive the Benefits of the Daily Regulation Activities, and How Satisfied Are Teachers at the End of the Day? To What Extent Are There Differences Among the Schools? (Question 2a)

The results showed that the day's activities were particularly perceived as beneficial for student learning and support of students, followed by beneficial for teachers but at almost a half standard deviation lower (see Table 12.10). The lowest were the

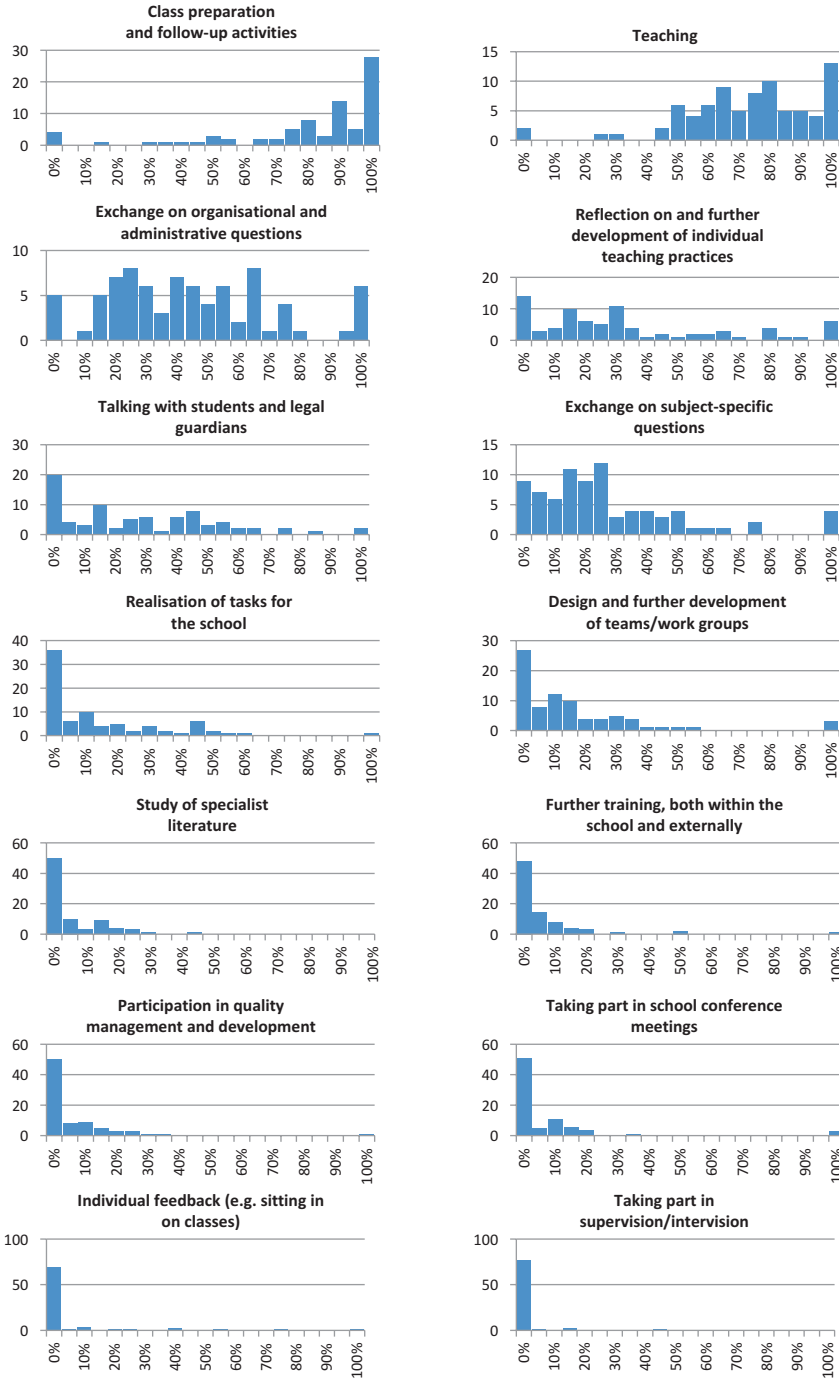


Fig. 12.2 Relative frequencies of different activities on a personal level
 Number of persons by average frequency of activities (summarized in levels of 5% each); 100% signifies that this activity was reported on each day an activity had been recorded; 0% signifies that it was not recorded on any of the documented days

Table 12.9 Average occurrence of different activities on a personal level by role (regulation activities shown in bold)

	Class teachers ^a n = 25	Subject teachers ^a n = 23	Teachers with leadership roles n = 30	Significant contrasts ^b
Group number	1	2	3	
Class preparation and follow-up activities	91.5%	78.1%	75.5%	1 > 3
Teaching	83.1%	68.6%	66.0%	1 > 2, 1 > 3
Exchange on organisational and administrative questions	54.3%	33.8%	42.0%	1 > 2
Reflection on and further development of individual teaching practices	40.6%	38.0%	20.6%	1 > 3, 2 > 3
Talking with students and legal guardians	37.0%	22.6%	24.3%	1 > 2, 1 > 3
Exchange on subject-specific questions	28.8%	25.3%	25.7%	Ns
Design and further development of teams/work groups	16.2%	16.1%	15.4%	Ns
Realisation of tasks for the school	10.8%	11.1%	18.7%	3 > 1
Taking part in school conference meetings	7.4%	10.2%	6.1%	Ns
Participating in quality management and development	1.9%	7.5%	9.2%	3 > 1
Further training, both within the school and externally	3.6%	9.3%	6.8%	Ns
Study of specialist literature	2.1%	7.3%	7.7%	2 > 1, 3 > 1
Individual feedback (e.g. sitting in on classes)	3.0%	6.5%	5.8%	Ns
Taking part in supervision/intervention	1.7%	0.5%	0.2%	Ns

Note. ^aGroups 'class teacher' and 'subject teacher' only comprise teachers with no school-related leadership roles

^bStatistically tested using binary logistic multilevel analyses on the level of daily activity entries. Contrasts with $p < .05$ were accounted for without an adjustment using the Holm-Bonferroni method

Table 12.10 Average perception of different forms of benefit and levels of satisfaction regarding the activities on a single day

Perceived benefit for... ^a	n	M	SD
Reaching educational objectives of students	899	6.8	2.0
Encouragement and support of students	897	6.9	1.9
Improvement/development of individual competencies	895	6.0	2.1
Improvement/development of individual teaching practices	897	6.0	2.1
Improvement/development of work done in teams	899	5.2	2.5
Improvement/development of the school as a whole	897	5.2	2.5
Level of satisfaction ^b	904	7.4	1.7

Data basis: daily entries regarding productivity perceptions and level of satisfaction (N = 947)

Note. ^aScale: 1 (not at all beneficial) to 10 (highly beneficial)

^bScale: 1 (not at all satisfied) to 10 (highly satisfied)

Table 12.11 Teachers' ratings of different forms of benefit and levels of satisfaction with the activities, by school

Perceived benefit for...	School 1 n = 254 M (SD)	School 2 n = 122 M (SD)	School 3 n = 229 M (SD)	School 4 n = 295 M (SD)	Significant contrasts ^c
Reaching educational objectives of students ^a	6.5 (2.2)	8.0 (1.6)	6.6 (2.1)	6.9 (1.7)	B > A, C, D
Encouragement and support of students	6.6 (2.1)	8.1 (1.5)	6.7 (2.0)	6.8 (1.7)	B > A, C, D
Improvement/development of individual competencies	6.1 (2.0)	6.7 (2.2)	5.7 (2.1)	5.9 (2.0)	–
Improvement/development of individual teaching practices	6.1 (1.9)	6.8 (2.1)	5.6 (2.2)	5.9 (1.9)	–
Improvement/development of work done in teams	4.9 (2.6)	6.0 (2.4)	5.1 (2.6)	5.2 (2.2)	–
Improvement/development of the school as a whole	5.1 (2.6)	6.1 (2.5)	5.1 (2.6)	4.9 (2.2)	–
Level of satisfaction regarding a single day ^b	7.4 (1.5)	8.0 (1.4)	7.4 (1.9)	7.1 (1.6)	–

Data basis: daily entries regarding the perceived benefit and level of satisfaction ($N = 947$)

Note. ^aScale: 1 (not at all beneficial) to 10 (highly beneficial)

^bScale: 1 (not at all satisfied) to 10 (highly satisfied)

^cStatistically tested using linear multilevel analyses (level 1: daily benefit/satisfaction; level 2: persons). Listed are contrasts with $p < .05$ with adjustment using the Holm-Bonferroni method

perceptions of benefit for developments on the team and school levels. The average level of teachers' daily satisfaction was rather high, with a mean of 7.4. Interestingly, the standard deviation was low.

If the average benefit ratings were calculated separately by schools, one school (school 2) would exhibit clear upward deviations (see Table 12.11). For the two benefit perceptions concerning students, the difference in relation to the other schools proved to be statistically significant, even with a correction of the multiple comparisons problem. Moreover, school 2 exhibited the highest levels of satisfaction for the survey period. However, after adjustment using the Holm-Bonferroni method, this difference was no longer significant. In contrast to the occurrence of activities (see Sect. 12.6.1.3 above), certain benefit ratings seemed to vary significantly between the schools, although it was only one school out of four that differed. Therefore, this result needs to be corroborated in a larger sample.

12.6.2.2 To What Extent Are Teachers' Daily Regulation Activities Related to Teachers' Daily Perceptions of Benefit and Teachers' Daily Satisfaction Levels? (Question 2b)

To answer this research question, the six statements concerning perceived benefit, based on factor analyses and high correlations within each factor, were combined into three learning and development-related benefit aspects, based on the object of

benefit: For the students, for the teachers, and for the team and the school. As Table 12.12 shows, the daily benefit rating for the students' learning process was positively associated with *teaching*, *class preparation and follow-up activities*, and *talking with students and legal guardians*, most of all. If the focus was on regulation activities, however, only less distinct connections appeared. *Reflection on and development of individual teaching practices* seemed to be positively related to teachers' daily benefit rating for student learning.

Overall, taking part in *further training, both within the school and externally* correlated in a slightly negative manner with teachers' perceived benefit for the students. As a consequence, further training was regarded as something from which the main target group was not able to benefit directly and as something that might even diminish the benefit, respectively.

Apart from that, *further training, both within the school and externally* was associated with the perceived benefit for the teachers themselves in a positive manner, together with *reflection on and further development of individual teaching practices* and *teaching*. The other statistically significant correlations with the development of the teachers were very low ($|r| < .10$, i.e. less than 1% explained variation).

Subsequently, perceived benefit for team and school development was related systematically but not very closely to numerous forms of activities in a positive manner, most of all *exchange on organisational and administrative questions* and discussion on the *design and further development of teams and work groups*. *Exchange on subject-specific questions*, *taking part in school conference meetings*, *participation in quality management and development*, *realisation of tasks for the school*, and *reflection on and further development of individual teaching practices* also correlated positively (in decreasing order). *Individual feedback (e.g. sitting in on classes)* was associated in a positive manner significantly as well, yet correlation strength was so low ($|r| < .10$, i.e. less than 1% explained variation) that this relation bears no meaning.

Further, there was no clear correlation between the recorded activities and the daily recorded *level of satisfaction*. Although two of the coefficients were significant ($p < .05$) – namely, *teaching* and *reflection on and further development of individual teaching practices* – correlation strength was below $|r| = .10$ or $r^2 = 1\%$ and, therefore, irrelevant. For this reason, the somewhat surprising negative significance of the correlation with *reflection on and further development of individual teaching practices* bears no meaning.

12.6.2.3 To What Extent Is Teachers' Perceived Daily Benefit Related to Their Daily Level of Satisfaction? To What Extent Do the Relations Between Daily Benefit and Satisfaction Differ Among the Schools? (Question 2c)

To answer this question, bivariate correlations between teachers' daily perceived benefit and daily level of satisfaction were calculated. Table 12.13 documents the Pearson correlation coefficients in general as well as separately for each school.

Table 12.12 Correlations between daily activities and different benefit ratings and level of satisfaction regarding the respective day (regulation activities shown in bold)

	Benefit for students n = 899	Benefit for teachers n = 898	Benefit for team/school n = 899	Satisfaction n = 904
Teaching	0.46***	0.18***	0.14***	0.09
Class preparation and follow-up activities	0.22***	0.06	-0.09*	0.00
Reflection on and further development of individual teaching practices	0.11**	0.25***	0.13***	-0.07
Exchange on organisational and administrative questions	0.07	0.02	0.31***	-0.04
Talking with students and legal guardians	0.20***	0.08	0.15***	0.05
Exchange on subject-specific questions	0.03	0.07	0.23***	-0.01
Design and further development of teams/work groups	-0.01	0.01	0.31***	-0.01
Participating in quality management and development	0.03	0.00	0.17***	0.03
Taking part in school conference meetings	0.00	0.04	0.19***	-0.01
Realisation of tasks for the school	-0.01	-0.06	0.17***	-0.02
Further training, both within the school and externally	-0.17***	0.16***	0.06	-0.01
Study of specialist literature	0.01	0.09	0.04	0.03
Individual feedback (e.g. sitting in on classes)	-0.04	-0.02	0.08	0.00
Taking part in supervision/intervision	0.03	0.05	0.05	0.03

Note. Data basis: daily entries (N = 947)

Pearson correlation coefficients. * $p < .05$, ** $p < .01$, *** $p < .001$ (with adjustment using the Holm-Bonferroni method for 14 relations at a time)

Again, the six statements concerning perceived benefit were combined into the three learning and development-related benefit aspects: Students, teachers, and team/school.

The results showed that teachers' daily level of satisfaction was related more closely to teachers' daily perceived benefits for student learning ($r = 0.38$, $p < .001$) and for the development of the teachers ($r = 0.34$, $p < .001$) than for team or school ($r = 0.15$, $p < .05$). Accordingly, the results revealed a higher importance of the perceived benefit for students and teachers than of the perceived benefit for the team and the school for teacher's individual daily satisfaction.

The four columns on the right side of Table 12.13 reflect the correlation strengths, separated by school and the multivariate calculations of R^2 for all three predictors (students, teachers, and team/school). None of the schools differed significantly.

Table 12.13 Correlations between teachers' daily perceived benefit and teachers' daily level of satisfaction

Correlation ^a between perceived benefit for different groups and level of satisfaction	Generally n = 897	School 1 n = 252	School 2 n = 121	School 3 n = 229	School 4 n = 295	p ^c
Students ^b	0.38***	0.35***	0.50***	0.41**	0.27*	ns
Teachers ^b	0.34***	0.33***	0.41**	0.35***	0.26*	ns
Team and school ^b	0.15*	0.17**	0.19 ns	0.13 ns	0.08 ns	ns
R squared (multivariate)	17.4%	17.0%*	28.5%*	19.0%*	10.6% ns	ns

Note. Data basis: daily entries (N = 947)

* p < .05, ** p < .01, *** p < .001

^aCalculation of bivariate correlation coefficients and multivariate variance explanation of the complete model in Mplus with standard errors corrected for the design effect (type = complex)

^bCombination of the two ratings of benefit for students, the teachers, and the team and the school by means of averaging at a time (based on a highly plausible three-dimensional factorial structure and reliability coefficients of alpha \geq 0.85)

^cStatistical testing by hierarchical linear regression with effects of school dummy variables (level 2) on the random slope of the effect of teachers' perceived daily benefit on daily satisfaction (level 1) (adjusted using the Holm-Bonferroni method)

Noteworthy, however, is that a deviation from the general tendency was found at two schools. Whereas teachers' daily level of satisfaction at school 2 appeared to be influenced by teachers' perceived benefit in an above-average manner with a total of approximately 28.5%, the explained variance at school 4 was lower and below average with 10.6%. It seems that at school 4, teachers' satisfaction was less dependent on the perceived benefit of their daily work. Instead, for teachers' perceived daily satisfaction at school 4, other factors may have been more influential (e.g. relationship with students, or with colleagues).

12.6.2.4 To What Extent Do Individual Factors Influence the Relation Between Teachers' Perceived Daily Benefit and Teachers' Daily Satisfaction Level? (Question 2d)

The analyses in Table 12.14 show if and to what extent individual factors were able to explain the variation in the correlation between daily perceived benefit and daily level of satisfaction. The analyses were conducted as a series of multilevel models, in which the correlation between teachers' perceived daily benefit and teachers' daily level of satisfaction was assessed on a personal level as a random slope. To explain the variation in the slopes, teachers' personal traits (sex, length of service, internal search interest, external search interest) were used as predictors.

There were no significant moderating effects for either teachers' *sex* or *length of service*. In contrast, there were rather distinct moderating effects for the teachers' internal search interest (having interest in knowledge concerning teaching quality and student learning) and external search interest (being open and ready to learn

Table 12.14 Influences of different individual factors on relation (random slope) between teachers’ perceived daily benefit for different areas and teachers’ daily level of satisfaction

Moderators for the linear effect of perceived daily benefit on daily level of satisfaction	Mean random slope (standard.)	b (on random slope)	se b	p	r ² (of random slope)
<i>Daily level of satisfaction regressed on perceived daily benefit for students</i>					
Sex (f = 1, m = 2)	.26***	-0.059	0.084	ns	0.7%
Length of service (in years)	.26***	0.005	0.005	ns	1.1%
Internal search interest	.26***	0.194	0.097	p < .05	4.4%
External search interest	.26***	0.248	0.092	p < .01	7.7%
<i>Daily level of satisfaction regressed on perceived daily benefit for teachers</i>					
Sex (f = 1, m = 2)	.23***	-0.016	0.088	ns	0.6%
Length of service (in years)	.23***	0.000	0.005	ns	0.6%
Internal search interest	.23***	0.084	0.091	ns	1.2%
External search interest	.23***	0.094	0.088	ns	1.5%
<i>Daily level of satisfaction regressed on perceived daily benefit for team and school</i>					
Sex (f = 1, m = 2)	.11***	-0.026	0.064	ns	1.1%
Length of service (in years)	.11***	-0.003	0.004	ns	2.2%
Internal search interest	.11***	0.184	0.062	p < .001	17.4%
External search interest	.11***	0.173	0.063	p < .01	13.9%

Note. Data basis: daily entries (N = 947) for benefit perceptions and for levels of satisfaction as well as for personal traits documented in the initial survey (N = 81)

Each line represents a separate multilevel model for a single moderator. The effects shown in column 3 are unstandardized regression coefficients of the level-2 moderator in column 1 on the random slope of the daily level of satisfaction regressed on the perceived daily benefit for different areas, both on level 1. *** p < .001

from others). For teachers that were interested in optimizing their practices, their daily work-related level of satisfaction depended more strongly on their perceived daily benefit than it did for teachers with less interest. However, this applied only regarding the benefit for student learning as well as for the teams and the school but not regarding benefit for the teachers.

12.7 Discussion

In this contribution, a newly developed time sampling-based method of assessing teachers’ daily regulation activities at secondary schools was explored empirically. For this purpose, in a first step, we developed a theoretical framework model, in which regulation in the context of school improvement is conceptualized by combining (self-)regulatory approaches from organization and school development research and pedagogical psychology. Accordingly, regulation of school-related activities is understood as the (self-)reflective individual, interpersonal, and organizational identification, analysis, and adaptation of tasks, dispositions, operations, and standards and goals by applying cognitive, metacognitive,

motivational-emotional, and resource-related strategies. Regulation means to reconstruct and deconstruct current practices and to further develop current practices by seeking new knowledge.

In a second step, a mixed-method case study was conducted at four secondary schools (in Switzerland) to identify teachers' regulation activities. We aimed to detect teachers' perceptions of the benefit of regulation activities for student learning and support of students, for the development of teaching competencies, and for the development of teams and schools. We focused on two sets of investigations: (1) analysis of the frequency of teachers' daily regulation activities at secondary schools and identifying differences between parts of the week, teachers, and schools, and (2) assessment of teachers' perceived benefit of the daily regulation activities and teachers' satisfaction and the relations between teachers' daily regulation activities, perceived daily benefit for different potential benefits, and daily levels of satisfaction. The results of both sets of questions were factored in for the assessment of the validity of the newly developed approach for daily measurement of teachers' regulation activities. Data analyses were based on 947 daily log entries of 81 teachers in total. Because of the high response rate in general and for each school, no severe systematic biases were expected. However, the sample size on the personal level has to be considered as rather small.

In summary, we found the following results for the *first of set of questions*: In accordance with the first hypothesis, (H1), teachers' most frequent regulation activities were found to be in the area of administration and organisation and in reflection on individual teaching practices. On average, the teachers reported these activities 1–2 times a week. Their average frequency is therefore relatively limited. Exchange with others on subject-related questions took place on only about 2 out of 10 days. Activities pertaining to team and school development appeared even less frequently, as did also regulation activities that require more introspection and initiative (e.g. intervention).

Teachers used the weekends basically for class preparation and follow-up activities. To a minor degree, the teachers used the weekend for reflection on and further development of their teaching practices and for exchange on organisational and administrative questions. We found plausible differences between teachers' activities during the week and activities on the weekend (e.g. teaching classes, exchange, reflection on individual teaching practices) as well as similarities (e.g. class preparation and follow-up activities) that are in line with previous research (H2). However, contrary to our expectations, teachers did not read specialist literature significantly more often on weekend days than on weekdays, although there was a slightly higher frequency on the weekend, as expected. This not significant result might be due to the very low level of regulation activity identified during the 3 weeks (study of specialist literature made up only 6% [$n = 60$] of the activities reported). Therefore, an extension of the data collection over a longer time (not only for 3 weeks) would perhaps help to elaborate this point more clearly. This could be useful as well for the analyses of other activities with a low occurrence during the 3 weeks (e.g. individual feedback).

In line with previous research, only random differences in the frequency of regulation activities appeared between schools (H3), in contrast to significant differences between teachers (H4) (Camburn & Won Han, 2017; Sebastian et al., 2017). These individual differences can be partly explained by the specific roles that the teachers have at the school (Pedder, 2007). As expected, teachers with leadership roles engaged more often in activities regarding school quality management and school development as well as in tasks for the school than teachers with no leadership roles did. Teachers with leadership roles reflected on their individual teaching practices less often and did not develop these further as often as class or subject teachers did, which was expected according to H5. That these differences were no longer significant when correcting for the alpha inflation problem, could be explained by the fact that teachers with leadership roles also teach classes. In Switzerland, therefore, the two groups are not distinct and may share more activities than is the case in countries where school leaders do not have to teach. Nevertheless, further studies should examine this aspect in more depth and in a larger sample.

The *second set of questions* assessed teachers' perceived benefit of the daily activities as well as teachers' daily satisfaction. As expected according to H6, the results revealed that teachers rated the regulation activities as especially beneficial for teaching, student learning, and teachers' learning but as less beneficial for team and school development. This is not surprising, since teacher education and professional development courses focus, above all, on teacher competencies in their core work area – that is, teaching. Additionally, 80% of the teachers' working hours were dedicated to teaching and fostering student learning. The lower level of perceived benefit for team and school development could be an indication that there is still need for support of activities in that area (Camburn & Won Han, 2017; Creemers & Kyriakides, 2012; Gutierrez, 2015).

As expected according to H7, teachers' perceived benefit of these activities varied school-specifically, although it was only one school (school 2) that outperformed the other three schools. Besides the need to corroborate this result in a larger sample, it will be crucial to work out to what extent school 2, at which the teachers rated the benefit for student learning and support of students as higher, differs from the other schools in other features (on the individual and school level). It could be that there was a stronger standard implemented at this school for teaching and the achievement of learning goals or professional competencies, and teachers' interest in reflection on school practices could differ from other schools in a positive manner. Taking into account the quantitative questionnaire survey data will make it possible to test these assumptions.

The results regarding correlation between daily regulation activities, daily perceived benefits, and daily levels of satisfaction partially confirm the hypotheses. In line with our assumption H8a, there was a positive, albeit weak, correlation between the activities that include reflection on and further development of individual teaching practices and teachers' ratings of the benefit for student learning. Further training, however, related negatively to teachers' perceived benefit for student learning. In light of the high demands placed on further training programmes in order to be effective for student learning, this result may be understandable (Day, 1999;

Desimone, 2009). However, further training as well as reflection on and further development of individual teaching practices were positively correlated with perceived benefits for the teachers themselves. As previous studies have shown, further training has an impact first of all on teachers' practices and beliefs, and only in second place, and under specific conditions, on student learning (Kreis & Staub, 2009).

Other regulation activities, however, seem to be connected only to the perceived benefit for team and school development but not for students and teachers, most of all exchange on organisational and administrative questions and further development of teams. The fact that more frequent exchange on subject-specific questions was, unexpectedly, not associated with higher levels of perceived benefit for the teachers themselves indicates that these activities are seen more as a service for the team and school than as a source of individual professional development. This means either that the quality of exchange has to be increased (see Spillane, Min Kim, & Frank, 2012, for the preconditions of effective exchange) or that the value and necessity of this important type of shared activity for professional development have to be made more visible.

Overall, the level of the correlations between the daily regulation activities and the thematically corresponding perceived benefits is somewhat lower than we would have expected. There are two possible explanations for this: First, the occurrence of an activity, e.g. exchange on subject-specific questions, may vary considerably in estimated quality and productivity. Activities perceived as unproductive will lower the correlation between the occurrence of activities and the perceived benefit. Second, the activities were unspecified not only regarding their perceived quality but also regarding the duration. By looking only at daily occurrences of activities (yes/no), very short sequences are treated in the same way as long ones, which also leads to lower correlations between activities and perceived benefits.

Our hypothesis H8b on the relation between teachers' daily regulation activities and teachers' daily level of satisfaction could be confirmed only partially. We expected that daily regulation activities are related systematically but on a weak level to teachers' daily level of satisfaction. However, the identified correlations were insignificant. Therefore, the occurrence of the regulation activities in itself had no effects on teachers' daily level of satisfaction. Instead, as argued in H8 and H9, the perceived benefits of the regulation activities are significantly related to the daily satisfaction level. Accordingly, and in line with school improvement and school effectiveness research (Creemers & Kyriakides, 2008; Hallinger & Heck, 2010) and self-regulated learning research (Wirth & Leutner, 2008), high-quality activities are more important for teachers' daily satisfaction than the quantity of the respective activities is. In line with H9, the strongest contribution to a high daily satisfaction level comes from teachers' perception that the daily activities are beneficial for student learning and for teachers' professionalisation and development of teaching practice (Landert, 2014). The more positive the perceived benefit, the more satisfied the teachers are at the end of the day.

For the question as to what extent the relation between daily benefit and daily satisfaction differ among the schools (H10), the results were similar to those for the

analysis for H7. The daily satisfaction levels at school 2 seemed to be influenced by the perceived benefit to a greater degree than at other schools; however, the effect was not significant. It may be that a larger sample providing more power would yield a different result.

The concluding moderator analyses showed, as expected according to H11, that it is plausible in general to assume that interest in searching for new knowledge (Mitchell & Sackney, 2011) has an effect on the relation between perceived benefit and satisfaction level. Teachers, who strive to do a better, more professional job by seeking to acquire more knowledge, appear to be more influenced in their perceptions of satisfaction by their perceived daily benefits than teachers with lower interest are. The results revealed this interaction to be especially relevant for achieving team and school development goals and, in a weakened form, for student learning.

Interestingly and against expectations, there was no significant moderation effect of interest in seeking new knowledge concerning further development of one's own teaching practices and competencies. The question arises as to how this result can be interpreted. As the mean level of perceived benefit for the teachers themselves and its standard deviation (Table 12.10) as well as the general association between this benefit (for teachers) and perceived daily satisfaction (Tables 12.13 and 12.14) are inconspicuous (since the correlation was between the coefficients for the benefit for students and for team and school), there are no technical reasons, such as restricted variance, for this lower level of moderation effect. Therefore, we exclude an artefact and, instead, try to find a content-specific interpretation.

A first possible explanation relates to the meaning of the moderators at issue – that is, internal interest and external interest in seeking new knowledge. Based on the operationalization applied, the two scales measure teachers' interest in monitoring the effectiveness of their own teaching for student learning and interest in seeking new knowledge for optimizing teaching and student learning. Our assumption is that not all of the assessed benefits are equally sensitive to these interests, and that these indicators of interest may not be equally interpreted as reflecting the actual value (Eccles & Wigfield, 2002) of the respective benefits. For instance, teachers may see the goals of this search for knowledge more in optimization of student learning and of team and school and not so much in further development of their own competencies. Daily activities that are perceived as productive for one's own person and one's own teaching may possibly for this reason per se contribute to teachers' daily satisfaction – namely, largely independently of teachers' interest in monitoring effectiveness and searching for new knowledge. However, for student learning and development of the team and school, interest in seeking new knowledge increases the importance of the daily activities for teachers' satisfaction, as is supposed by expectancy-value theory (Eccles & Wigfield, 2002). If this explanation were correct, it would be helpful in the future to assess the benefit of such activities not only indirectly via teachers' interest but also directly.

The particularly strong moderation effect in connection with benefit for team and school could be related to the fact that precisely the mean association between perceived benefit for team and school development and satisfaction, in contrast to the other two areas of benefit, is definitely lower, at $r = .15$ (vs. $r = .34$ and $r = .38$). The

perceived benefit of team and school activities thus appears to contribute on average only little to teachers' satisfaction. According to Landert (2014), teachers' work satisfaction in Switzerland is based mainly on what are viewed as teachers' core activities – namely, teaching and supporting students. In contrast, team and school development activities are seen by teachers often as additional to their core mission and, moreover, as difficult and connected with stressful situations, such as the introduction of reforms. Unless they have specific interest in these activities, it appears that teachers benefit little from them for their own satisfaction.

A second possible explanation for the lack of a moderator effect could be that teachers view their own competencies as a relatively static given and not as plastic, malleable, and capable of development, as is the case for students or team. Following Dweck and Leggett (1988), then, teachers' implicit theories must differ depending on the learning object being focused on: Regarding their own competencies, teachers would have a more fixed mindset (as opposed to a growth mindset) and, thus, a belief that their own competencies are not or are only little modifiable, whereas their mindset regarding student learning or further development of the team or school would be more of a growth mindset. Fixed mindsets tend to lead to lower interest in further development of one's own competencies and also have a negative effect on the achievement of objectives. This supplementary hypothesis cannot be tested further based on the existing data, as in the present study, no information is available on those views and beliefs. Further studies will be needed to clarify the issue.

12.8 Strengths and Weaknesses of the Applied Methodological Approach, and the Need for Further Research

Considering the results, presented above, and the confirmation of most of the hypotheses, it can be concluded that the newly developed methodological approach makes an instrument available that appears to be suitable for recording teachers' daily regulation activities in a (relatively) valid manner and for use as a complement tool to existing instruments, such as standardized surveys for retrospective recording of regulation activities. Daily micro-level measurements, such as those employed in this study, are unique in uncovering differences between parts of the week, teachers, and (to some extent) schools, and this allows for the recording of individual as well as collective regulation activities profiles. Further, it is crucial in this context that the activities are recorded not only on a daily level but also for different areas. That means that information can be obtained on regulation activities for teaching or administrative/organisational matters as well as for team and school development. In addition, in the case study, school leaders and selected teachers confirmed in interviews, conducted after data collection, that the methods chosen, indeed, capture the main activity areas of the teachers with an appropriate degree of differentiation.

It became clear that the combination of recording the frequency of regulation activities and collecting information on the perceived daily benefit increased the substance of the results. Particularly the finding that it is not the realization of regulation activities but rather the perceived benefit of daily regulation activities that is systematically associated with perceived daily satisfaction confirms that it is necessary to capture not only the quantities but rather the qualities of activities (Creemers & Kyriakides, 2008).

However, precisely in that regard, there is a deficit in the design of the case study, insofar as perceived benefit was not rated for each individual activity but only at the end of the day as a kind of balance sheet. When planning the case study, we had intended to implement ratings for each activity. However, after intensive discussions with teachers, we had to drop that as we feared that for the teachers, benefit ratings of every single activity would have been a burden in terms of time (and also in part in terms of content). This would have been the case, especially for short activities, the benefit of which for different aspects would be difficult to determine. Based on the analyses, however, this must be reconsidered, particularly as from this a clearer and closer relation between regulation activities and perceived benefit is expected.

Further studies will also be necessary in order to include in the analyses not only daily frequencies but also the time spent on the individual activities within the day. Not yet considered in the findings presented here is also the social structure of the regulation activities – that is, whether teachers carried them out alone or together with others. We plan to include that aspect in further analyses.

A major limitation of the case study, presented here, is that we examined only four schools so that analysis of differences among schools was possible only to a limited extent. It, therefore, remains open whether or not schools differ in the frequency of regulation activities (Camburn & Won Han, 2017; Sebastian et al., 2017), also under consideration of more in-depth analysis, as is possible with time-sampling data. Regarding the quality of the regulation activities, we expected to find differences (H7), which the case study confirmed in part. However, the differences were only very small, so that it will also be necessary to check the results in a larger sample of schools.

A further limitation is that it was not possible to set teachers' regulation activities in overall relation to the concrete development of student learning, to teaching, or to school development. It remains to be seen whether or not these activities are not only subjectively but in fact verifiably beneficial to further development of a teacher's own competencies, of teaching, and of team and school. From a methodological perspective, it also remains an open question whether the data collected represent a better basis for explaining differences in student performance and student performance development. This is a relevant question, ultimately, also from an economic perspective because compared to filling in a standardized questionnaire, the effort that the data collection required of the teachers, even though it was not very great (5–10 minutes per day), should not be underestimated.

Beyond that, an important question concerning the validity of the methodological approach is the time point of data collection. The data were collected in 3 weeks during the second quarter of the school year, with each week being followed by a

week with no data collection. In contrast to the number of days on which data had to be collected in order to obtain a stable data base (Bolger, Stadler, & Laurenceau, 2012), there were practical considerations for the choice of these 3 data collection weeks and the on-off rhythm. For example, the data collection period could not be expanded to an entire school year, as it would then not be possible to provide each school with individual feedback within the same year. Ultimately, the procedure chosen could also limit the validity of the design and explain why certain regulation activities, such as further training or intervision, were seldom recorded. Whether this, in fact, corresponds to reality or whether a different frequency would be observed if we examined an entire school year, would have to be checked. In one interview with a school leader after data collection, we learned that the school conducted most of its internal further training programmes in the second half of the school year. With this, it can be assumed that precisely those regulation activities that are not normally carried out throughout the entire school year cannot be adequately represented using the methodological approach applied here. And, even though we found no indications for it based on the interviews that we conducted, the opposite is also conceivable – that in the study, certain regulation activities were identified more frequently than they appear in reality because in the data collection period, there was by chance a particular focus on, for instance, exchange and cooperation, and intensive exchange did not take place all throughout the year.

All in all, then, it will be important to conduct further analyses and to test the chosen methodological approach in further studies.

References

- Adams, E. L., Carrier, S. J., Minotue, J., Porter, S. R., McEachin, A., Walkowiak, T. A., & Zulli, R. A. (2017). The development and validation of the instructional practices log in science: A measure of K-5 science instruction. *International Journal of Science Education, 39*(3), 335–357.
- Altrichter, H., & Kemethofer, D. (2015). Does accountability pressure through school inspections promote school improvement? *School Effectiveness and School Improvement, 26*(1), 32–56.
- Anusic, I., Lucas, R. E., & Donnellan, M. B. (2016). The validity of the day reconstruction method in the German socio-economic panel study. *Social Indicators Research, 1*–20. <https://doi.org/10.1007/s11205-015-1172-6>
- Argyris, C., & Schön, D. (1996). *Organisational learning II: Theory, method and practice*. Reading, MA: Addison Wesley.
- Bolger, N., Stadler, G., & Laurenceau, J.-P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285–301). New York, NY/London, UK: Guilford Press.
- Butler, D. L., Novak Lauscher, H., Jarvis-Selinger, S., & Beckingham, B. (2004). Collaboration and self-regulation in teachers' professional development. *Teaching and Teacher Education, 20*, 435–455.
- Camburn, E. M. (2010). Embedded teacher learning opportunities as a site for reflective practice: An exploratory study. *American Journal of Education, 116*, 463–489. <https://doi.org/10.1086/653624>

- Camburn, E. M., Spillane, J. P., & Sebastian, J. (2010). Assessing the utility of a daily log for measuring principal leadership practice. *Educational Administration Quarterly*, 46(5), 707–737.
- Camburn, E. M., & Won Han, S. (2017). Teachers' professional learning experiences and their engagement in reflective practice: A replication study. *School Effectiveness and School Improvement*, 28(4), 527–554.
- Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational Evaluation and Policy Analysis*, 23(2), 145–170.
- Creemers, B. P. M., & Kyriakides, L. (2008). The dynamics of educational effectiveness. In *A contribution to policy, practice and theory in contemporary schools*. London, UK/New York, NY: Routledge.
- Creemers, B. P. M., & Kyriakides, L. (2012). *Improving quality in education. Dynamic approaches to school improvement*. New York, NY: Routledge.
- Day, C. (1999). *Continuing professional development*. London, UK: Falmer Press.
- Day, C., & Sachs, J. (Eds.). (2004). *International handbook on the continuing professional development of teachers*. Maidenhead, UK: Open University Press.
- Desimone, L. M. (2009). Improving impact studies of teacher's professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181–199.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256–273.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132.
- Elliott, S. N., Roach, A. T., & Kurz, A. (2014). Evaluating and advancing the effective teaching of special educators with a dynamic instructional practices portfolio. *Assessment for Effective Intervention*, 39(2), 83–98.
- Fend, H. (2006). *Neue Theorie der Schule. Einführung in das Verstehen von Bildungssystemen. Lehrbuch*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
- Fussangel, K., Rürup, M., & Gräsel, C. (2010). Lehrerfortbildung als Unterstützungssystem. In H. Altrichter & K. Maag Merki (Eds.), *Handbuch Neue Steuerung im Schulsystem* (pp. 327–354). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
- Glennie, E. J., Charles, K. J., & Rice, O. N. (2017). Teacher logs: A tool for gaining a comprehensive understanding of classroom practices. *Science Educator*, 25(2), 88–96.
- Gräsel, C., Fussangel, K., & Parchmann, I. (2006). Lerngemeinschaft in der Lehrerfortbildung. Kooperationserfahrungen und -überzeugungen von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9(4), 545–561.
- Gräsel, C., Fußangel, K., & Pröbstel, C. (2006). Lehrkräfte zur Kooperation anregen – eine Aufgabe für Sisyphos? *Zeitschrift für Pädagogik*, 52(6), 205–219.
- Gutierrez, S. B. (2015). Teachers' reflective practice in lesson study: A tool for improving instructional practice. *Alberta Journal of Educational Research*, 63(3), 314–328.
- Hadwin, A. F., Järvelä, S., & Miller, M. (2011). Self-regulated, co-regulated, and socially shared regulation of learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 65–84). New York, NY/Milton Park, UK: Routledge.
- Hallinger, P., & Heck, R. H. (2010). Collaborative leadership and school improvement: Understanding the impact on school capacity and student learning. *School Leadership & Management*, 30(2), 95–110.
- Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability*, 26(1), 5–28.
- Hopkins, D., Stringfield, S., Harris, A., Stoll, L., & Mackay, T. (2014). School and system improvement: A narrative state-of-the-art review. *School Effectiveness and School Improvement*, 25(2), 257–281.
- Järvelä, S., & Järvenoja, H. (2011). Socially constructed self-regulated learning and motivation regulation in collaborative learning groups. *Teachers College Record*, 113(2), 350–374.

- Järvelä, S., Volet, S., & Järvenoja, H. (2010). Research on motivation in collaborative learning: Moving beyond the cognitive-situative divide and combining individual and social processes. *Educational Psychologist, 45*(1), 15–27.
- Järvenoja, H., Järvelä, S., & Malmberg, J. (2015). Understanding regulated learning in situative and contextual frameworks. *Educational Psychologist, 50*(3), 204–219.
- Johnson, E. (2013). *The impact of instructional coaching on school improvement*. Available from ProQuest Dissertations & Theses A&I. (1427330426). Retrieved from <https://search.proquest.com/docview/1427330426?accountid=14796>
- Kreis, A., & Staub, F. (2009). Kollegiales Unterrichtscoaching. Ein Ansatz zur kooperativen und fachspezifischen Unterrichtsentwicklung im Kollegium. In K. Maag Merki (Ed.), *Kooperation und Netzwerkbildung. Strategien zur Qualitätsentwicklung in Schulen* (pp. 26–39). Seelze, Germany: Klett-Kallmeyer.
- Kreis, A., & Staub, F. (2011). Fachspezifisches Unterrichtscoaching im Praktikum. Eine quasi-experimentelle Interventionsstudie. *Zeitschrift für Erziehungswissenschaft, 14*(1), 61–83.
- Kurz, A., Elliott, S. N., Kettler, R. J., & Yel, N. (2014). Assessing students' opportunity to learn the intended curriculum using an online teacher log: Initial validity evidence. *Educational Assessment, 19*(3), 159–184.
- Kwakman, K. (2003). Factors affecting teachers' participation in professional learning activities. *Teaching and Teacher Education, 19*, 149–170.
- Kyndt, E., Gijbels, D., Grosemans, I., & Donche, V. (2016). Teachers' everyday professional development: Mapping informal learning activities, antecedents, and learning outcomes. *Review of Educational Research, 86*(4), 1111–1150.
- Landert, C. (2014). *Die Berufszufriedenheit der Deutschschweizer Lehrerinnen und Lehrer (2014). Bericht zur vierten Studie des Dachverbandes Lehrerinnen und Lehrer Schweiz (LCH)*. Zürich, Switzerland: Landert Brägger Partner.
- Lomos, C., Hofman, R. H., & Bosker, R. J. (2011). Professional communities and student achievement – A meta analysis. *School Effectiveness and School Improvement, 22*(2), 121–148.
- Louis, K. S., Kruse, S., & Marks, H. M. (1996). Schoolwide professional community. In F. M. Newmann (Ed.), *Authentic achievement. Restructuring schools for intellectual quality* (pp. 179–203). San Francisco, CA: Jossey-Bass Publishers.
- Meredith, C., Moolenaar, N. M., Struyve, C., Vandecandelaere, M., Gielen, S., & Kyndt, E. (2017). The measurement of collaborative culture in secondary schools: An informal subgroup approach. *Frontline Learning Research, 5*(2), 24–35.
- Messmann, G., & Mulder, R. H. (2018). Vocational education teachers' personal network at school as a resource for innovative work behaviour. *Journal of Workplace Learning, 30*(3), 174–185.
- Mitchell, C., & Sackney, L. (2009). *Sustainable improvement: Building learning communities that endure*. Rotterdam, The Netherlands: Sense Publishers.
- Mitchell, C., & Sackney, L. (2011). *Profound improvement. Building learning-community capacity on living-system principles* (2nd ed.). London, UK/New York, NY: Routledge.
- Muijs, D., Harris, A., Chapman, C., Stoll, L., & Russ, J. (2004). Improving schools in socio-economically disadvantaged areas – A review of research evidence. *School Effectiveness and School Improvement, 15*(2), 149–175.
- Nguyen, Q. D., Fernandez, N., Karsenti, T., & Charlin, B. (2014). What is reflection? A conceptual analysis of major definitions and a proposal of a five-component model. *Medical Education, 48*, 1176–1189.
- Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research. An introduction and some practical recommendations. *Journal of Personnel Psychology, 9*(2), 79–93. <https://doi.org/10.1027/1866-5888/a000009>
- Oude Groote Beverborg, A., Geerlings, J., Slegers, P. J. C., Feldhoff, T., van Veen, K., & Wijnants, M. (2017). *Diversity in learning trajectories. Towards a tangible conceptualization of dynamic processes*. Paper presented in an invited symposium. Paper presented at the 17th Biennial EARLI Conference, Tampere, Finland.

- Oude Groote Beverborg, A., Slegers, P. J. C., Endedijk, M. D., & van Veen, K. (2017). Towards sustaining levels of reflective learning: How do transformational leadership, task interdependence, and self-efficacy shape teacher learning in schools. In K. Leithwood, J. Sun, & K. Pollock (Eds.), *How school leaders contribute to student success* (Studies in educational leadership) (Vol. 23, pp. 93–129). Cham, Switzerland: Springer.
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Pedder, D. (2007). Profiling teachers' professional learning practices and values: Differences between and within schools. *The Curriculum Journal*, 18(3), 231–252.
- Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory Into Practice*, 41(4), 219–225.
- Raes, E., Boon, A., Kyndt, E., & Dochy, F. (2017). Exploring the occurrence of team learning behaviours in project teams over time. *Research Papers in Education*, 32(3), 376–401.
- Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222). New York, NY: Cambridge University Press.
- Schön, D. A. (1984). *The reflective practitioner: How professionals think in action*. New York, NY: Basic Books.
- Schweizerische Konferenz der kantonalen Erziehungsdirektoren. (1999). *Reglement über die Anerkennung von Hochschuldiplomen für Lehrkräfte der Vorschulstufe und der Primarstufe*. Bern.
- Sebastian, J., Comburn, E. M., & Spillane, J. P. (2017). Portraits of principal practice: Time allocation and social principal work. *Educational Administration Quarterly*, 54(1), 47–84.
- Spillane, J. P., & Hunt, B. R. (2010). Days of their lives: A mixed-methods, descriptive analysis of the men and women at work in the principal's office. *Journal of Curriculum Studies*, 42(3), 293–331.
- Spillane, J. P., Min Kim, C., & Frank, K. A. (2012). Instructional advice and information providing and receiving behavior in elementary schools: Exploring tie formation as a building block in social capital development. *American Educational Research Journal*, 49(6), 1112–1145. <https://doi.org/10.3102/0002831212459339>
- Spillane, J. P., & Zuberi, A. (2009). Designing and piloting a leadership daily practice log. Using logs to study the practice of leadership. *Educational Administration Quarterly*, 45(3), 375–423.
- Spörer, N., & Brunstein, J. C. (2006). Erfassung selbstregulierten Lernens mit Selbstberichtsverfahren: Ein Überblick zum Stand der Forschung. *Zeitschrift für Pädagogische Psychologie*, 20, 147–160.
- Stringfield, S., Reynolds, D., & Schaffer, E. C. (2008). Improving secondary students' academic achievement through a focus on reform reliability: 4- and 9-year findings from the high reliability schools project. *School Effectiveness and School Improvement*, 19(4), 409–428.
- Vangrieken, K., Meredith, C., Packer, T., & Kyndt, E. (2017). Teacher communities as a context for professional development: A systematic review. *Teacher and Teacher Education*, 61, 47–59.
- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 21, 1–19.
- Weick, K. E. (1995). *Sensemaking in organizations*. London, UK: Sage.
- Weick, K. E. (2001). *Making sense of the organization*. Malden, MA: Blackwell Publishing.
- West, L., & Staub, F. (2003). *Content-focused coaching (SM): Transforming mathematics lessons*. Portsmouth, NH: Heinemann.
- Widmann, A., Mulder, R. H., & Köning, C. (2018). Team learning behaviours as predictors of innovative work behaviour. A longitudinal study. *Innovation*. <https://doi.org/10.1080/14479338.2018.1530567>.
- Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, 45(4), 267–176.

- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in education theory and practice* (pp. 277–304). Mahwah, NJ: Lawrence Erlbaum Association.
- Winne, P. H., & Hadwin, A. F. (2010). Self-regulated learning and socio-cognitive theory. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (Vol. 5, pp. 503–508). Amsterdam, The Netherlands: Elsevier.
- Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence. Implications of theoretical models for assessment methods. *Journal of Psychology*, 216(2), 102–110.
- Wolters, C. A. (2003). Regulation of motivation. Evaluating an underemphasized aspect of self-regulated learning. *Educational Psychologist*, 38(4), 189–205.
- Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 1–37). Mahwah, NJ: Lawrence Erlbaum Association.
- Zimmerman, B. J., & Schunk, D. H. (Eds.). (2001). *Self-regulated learning and academic achievement: Theoretical perspectives*. Mahwah, NJ: Lawrence Erlbaum Association.
- Zimmerman, B. J., & Schunk, D. H. (Eds.). (2007). *Motivation and self-regulated learning: Theory, research, and applications*. Mahwah, NJ/London, UK: Lawrence Erlbaum Association.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 13

Concept and Design Developments in School Improvement Research: General Discussion and Outlook for Further Research



**Tobias Feldhoff, Katharina Maag Merki, Arnoud Oude Groote Beverborg,
and Falk Radisch**

This book aimed to present innovative designs, measurement instruments, and analysis methods by way of illustrative studies. Through these methodology and design developments, the complexity of school improvement in the context of new governance and accountability measures can be better depicted in future research projects. In this concluding chapter, we discuss what strengths the presented methodologies and designs have and to what extent they do better justice to the multilevel, complex, and dynamic nature of school improvement than previous approaches. In addition, we outline some needs for future research in order to gain new perspectives for future studies.

In this discussion we are guided by Feldhoff and Radisch's framework on complexity (see Chap. 2). The chapters in this volume contribute in particular to discussion of the following aspects:

- The longitudinal nature of the school improvement process
- School improvement as a multilevel phenomenon
- Indirect and reciprocal effects
- Variety of meaningful factors

T. Feldhoff (✉)
Johannes Gutenberg University, Mainz, Germany
e-mail: feldhoff@uni-mainz.de

K. Maag Merki
University of Zurich, Zurich, Switzerland

A. Oude Groote Beverborg
Radboud University Nijmegen, Nijmegen, The Netherlands

F. Radisch
University of Rostock, Rostock, Germany

© The Author(s) 2021

A. Oude Groote Beverborg et al. (eds.), *Concept and Design Developments in School Improvement Research*, Accountability and Educational Improvement, https://doi.org/10.1007/978-3-030-69345-9_13

303

13.1 The Longitudinal Nature of the School Improvement Process

Even though school improvement always implies a change (Stoll & Fink, 1996), studying school improvement longitudinally was surprisingly neglected for a long time (Feldhoff, Radisch, & Klieme, 2014). For this reason, it is particularly important that four of the contributions in this volume (Chaps. 9, 10, 11, and 12) examine school improvement processes longitudinally. All of them use logs as a measurement instrument. Three of them use logs to capture microprocesses. The chapters show that logs can be used both in open form for qualitative analyses and in standardized form for quantitative analyses.

The chapters demonstrate several advantages of logs. Logs have the potential to capture day-to-day behaviour in the context of school improvement, and it is precisely in that area that there is currently a lack of established instruments. Day-to-day behaviour (and other microprocesses) cannot be captured using most traditional questionnaires, because they were developed for cross-sectional designs. Moreover, qualitative studies seldom apply a methodology designed to carefully examine microprocesses longitudinally.

Logs have the advantage of having higher validity than traditional questionnaires that focus more on the measurement of abstracted activities from a longer period of time (Anusic, Lucas, & Donnellan, 2016; Ohly, Sonnentag, Niessen, & Zapf, 2010; Reis & Gable, 2000). Logs can provide better insights into day-to-day activities and their dynamics. This means that also shorter time periods and shorter intervals between the measurements can be examined. Both play an important role in investigation of the highly dynamic and very diverse school improvement processes frequently found in schools, such as initiation of changes, team building, the handling of pressing problems, and so on.

Exactly these processes must be investigated, if the aim is to better understand school improvement in the context of new governance and accountability measures. Data gathered with standardized logs can be analyzed using many established statistical methods for time series analysis (Hamaker, Kuiper, & Grasman, 2015; McArdle, 2009; Valsiner, Molenaar, Lyra, & Chaudhary, 2009). Furthermore, with sufficiently large samples and measurement points, logs allow multilevel analysis and thus the analysis of interaction effects between the different levels, such as between school, person, and time. One methodology that is particularly geared towards processes and dynamics of individuals, as presented by Oude Groote Beverborg et al. (Chap. 11), allows the analysis of regularity and stability of (the coupling between) microprocesses and improvement. Using qualitative logs that were sensitive to local and personal circumstances and Recurrence Quantification Analysis, they were able to analyze the extent to which differences in the regularity and frequency of teacher reflection in the context of workplace learning are connected with their own developments.

The more qualitative methodologies presented in this volume (Chaps. 9, 10, and 11) also allow to acquire more detailed findings on the extent to which attitudes,

orientations, and perspective towards school tasks and school improvement processes change. However, the particular challenge these kinds of studies face is the identification of substantial changes and to differentiate them from more random or insignificant developments. Therefore, the illustrative studies' log-based methodologies, as well as the corresponding conceptualizations and theories, need to be further developed and applied to different situations and school improvement contexts. This is particularly relevant in connection with questions pertaining to new governance and accountability measures. Previous research has insufficiently studied how teachers and school leaders, as well as other actors, react to external demands or monitoring outcomes, integrate them in their school practices (or not), and utilize them for teaching and student learning (or not). Commonly used questionnaires or interviews capture retrospective self-reports and are thus limited in tapping into ongoing improvement processes. In this regard, the methodological and theoretical developments presented in Chaps. 9, 10, 11, and 12 hold the promise of a substantial gain in knowledge and a significant broadening and deepening of understanding the connection between accountability and school improvement.

A prerequisite for the use of logs to capture behaviour in a day-to-day manner is the validity of the log itself. How logs can be validated ideally using observations and interviews is described in the contribution by Spillane and Zuberi (Chap. 9). Beyond that, there are additional challenges that must be tackled, because of the temporal nature of change and development in school practices, the role of actors' motivations or perspectives within school improvement processes, or monitoring procedures. A main keyword here is 'measurement invariance.' The contributions by Lomos (Chap. 4) and Sauerwein and Theis (Chap. 5) provide insight into analyses for testing measurement invariance using Multiple Group Confirmatory Factor Analysis (MGCFA). Although the analyses presented in these two contributions are based on cross-sectional data, MGCFA can be used to assess whether the meaning of a construct remains stable across different time points. In addition, MGCFA allows the examination of change in understanding of a construct itself or differences between groups in their (change of) understandings of a construct.

Especially regarding the interpretation of findings on measurement invariance (or measurement variance), however, there are a number of substantial research gaps. Measurement (in)variance can be technically determined, but the interpretation of such a finding depends on one's theory. A finding that points to measurement variance could – from a methodological viewpoint – indicate that longitudinal analysis should not be conducted. However, the finding could also indicate that the meaning of the items within a construct has changed over time for the participants. This is often the very goal of a school improvement measures, for instance, when the aim is to implement collegial cooperation or raise commitment. In the future, therefore, findings should be carefully considered on their methodological and theoretical merit, and separated using suitable methodologies when needed.

Also needed are measurement instruments that are specifically developed for empirically depicting the developmental courses of processes. This is particularly important for processes where development means not simply 'more of the same,' such as in the form of higher approval, intensity, and so on, but where the construct

itself changes. For example, with collegial cooperation, rudimentary cooperation is characterized simply by exchange of materials, whereas high-quality cooperation is characterized by co-constructive development of concepts and materials (Decuyper, Dochy, & Van den Bossche, 2010; Gräsel, Fußangel, & Pröbstel, 2006). Accordingly, forms of adaptive measurement could be developed in school improvement research, something that is being done for some time now in the area of competency assessment (Eggen, 2008; Meijer & Nering, 1999). Alternatively or concomitantly, researchers could work together with practitioners in common contexts to co-develop scales and the meaning of their intervals.

13.2 School Improvement as a Multilevel Phenomenon: The Meaning of Context for School Improvement

School improvement processes make up a complex phenomenon that takes place at different levels not only within the education system but also within schools. Accordingly, the notion of ‘context’ is quite complex.

As discussed in the contribution by Reynolds and Neeleman (Chap. 3), the improvement of schools and the underlying processes depend heavily on the social, socioeconomic, and cultural context of the school, as well as on the accountability modus that is implemented in the particular education system. In this sense, context refers to political, cultural, and social factors external to the school. Within schools, however, the organization (e.g. leadership) might be the context for teachers’ team learning, and consequently, teachers’ team learning can be understood as a context for teachers’ learning and teaching.

In the last 20 years, many empirical studies have shown that it is essential to consider these nested structures at the appropriate levels when investigating school improvement processes (see Hallinger & Heck, 1998; Heck & Thomas, 2009; Van den Noortgate, Opdenakker, & Onghena, 2005). However, there are several problems and challenges, particularly regarding the analysis of the multilevel structure of school improvement and the issue of how different contexts can be identified and taken into account. Several chapters in this volume discuss these points in detail.

First of all, the chapters in this volume that used logs in order to investigate day-to-day activities (for example, the contributions by Spillane and Zuberi and by Maag Merki et al.) point out that in school improvement research the hierarchical structure must be extended to include (at least) two further levels: daily activities and individual activities. The level of daily activities can then be considered as ‘nested in persons’, and the individual activities are then activities ‘nested in days’. With this, an extensive nesting structure of school improvement processes unfolds: individual activities, nested in days, nested in persons, nested in teams, nested in schools, nested in districts or regions, nested in countries. Development of the appropriate methodology and empirical assessment of this structure is challenging and future school improvement research could concentrate on that.

To take account of the hierarchical structure, hierarchical multilevel analyses have become the standard (e.g. Luyten & Sammons, 2010). Nevertheless, Schudel and Maag Merki (Chap. 12 in this volume) have critically discussed the existing practice of multilevel analysis. Although nested structures are taken into account in multilevel analysis, for instance through correction of standard errors, important information is lost with the common aggregation of data (which allows the use of information at higher levels). In addition, current research focuses solely on the group mean as a measure for shared properties. Variances in the aggregated properties or other parameters in the composition of these properties are thus overlooked. Therefore, as Schudel and Maag Merki mention, multilevel models in educational research have to consider the double character of groups: global group properties emerge from the group level and group composition properties emerge from the lower, individual level. Moreover, educational researchers have to take into account the possibility of both shared properties and configural properties of group compositions. In this way, the composition of the teaching staff, as well as the position of the individual within the teaching staff, can be regarded as an independent and process-relevant aspect of the multilevel structure, and the relation of either or both with individual teacher's actions and experiences can be examined. The use of the Group Actor-Partner Interdependence Model (GAPIM) allows a more differentiated modelling of, for instance, the frequently observed divergence in actors' perspectives on the implementation of reforms or their divergence in handling accountability requirements (e.g. interested and motivated teachers versus those who are opposed). Thus, the GAPIM allows a more valid investigation of how school improvement measures affect teachers' instructions and students' learning.

Further questions that could be interesting for both school improvement research and assessment of accountability processes are, for example: What dynamics emerge out of which (properties) of group compositions? What changes in composition are affected by school improvement measures (such as measures to develop a shared educational understanding, to reach an agreement on guiding principles, and so on)? Can different developmental courses in schools be explained by group composition properties? What aspects of the composition of the teaching staff are important for the success of school improvement measures?

Ng (Chap. 7) argued for another approach to identifying school-internal context conditions: social network analysis. This methodology has only been adopted in a few studies up to now (Moolenaar, Slegers, & Daly, 2012; Spillane, Hopkins, & Sweet, 2015; Spillane, Shirrell, & Adhikari, 2018). Social network analysis allows examination of the social structure of school teams and investigation of how this structure affects teachers' practices and the school's improvement processes. A clear gain over other methodologies is that the loosely coupled structures of schools (Weick, 1976) can be made visible. As such, formal and informal team structures, as well as densities of ties within teams and with other actors, can be investigated with respect to sustainable school improvement. In addition, the methodology also makes it possible to compare individual schools, which may uncover explanations for school-specific developmental trajectories of students.

Vanblaere and Devos (Chap. 10) investigated the effect of context from yet another perspective. Their focus was on a school-specific innovation, which they assessed with qualitative teacher logs over the course of a year in four primary schools, which were characterized as either a high or a low professional learning community (PLC). With such qualitative logs, it is possible to assess developments in each separate school, while taking different starting conditions (low and high PLC) into account. When using such unstandardized logs, developmental courses and events can be captured that had not been anticipated in advance.

The presented studies open up new perspectives to include context in the study of school improvement and school practices. However, many aspects are still not taken into sufficient consideration. In particular, investigations of how aspects of contexts affect actors should be extended with detailed assessments of the extent to which actors themselves change their contexts through their perceptions of, and actions in, those contexts (Giddens, 1984). This continuous interaction would require a longitudinal design and methodology in addition to multilevel methodology, and this has not been considered enough in previous research. Measurement instruments must therefore be sufficiently sensitive regarding differences in contexts but also regarding the identification of changes (at different levels), which is a double challenge. Beyond that, more differentiated investigation is needed on the extent to which school improvement strategies are dependent on certain contexts to be functional for sustainable development, or on what strategies are particularly productive for schools with either high or low school improvement capacities. This raises the issue of generic or specific school improvement processes and success factors (Kyriakides, 2007).

13.3 Indirect and Reciprocal Effects

School improvement is a complex process in which many processes (e.g. leadership actions, decisions and actions of several teams, and individual teachers) are involved over time. This process takes place at different levels (school level, team level, classroom level). From this point of view, school improvement processes usually have direct and indirect effects. Twenty years ago, Hallinger and Heck (1998) already pointed out for school leadership research that ignoring indirect effects impacts the validity of findings on the effect of school principals' actions on student achievement. The same can be assumed also for school improvement processes and for processes connected with accountability requirements and reforms. Due to the number of factors involved in those processes and the resulting number of hypothetically possible direct and indirect effects, it is not possible to assess all direct and indirect effects simultaneously (for example using structural equation models). Here it is important to carefully consider what direct and indirect effects should be included in the theoretical and the empirical model, and, where needed, to test individual paths one after the other and in advance.

Indirect relations were addressed in the contribution by Ng (Chap. 7). Ng describes an example of a social network analysis that was used to identify heterarchical paths of decision-making processes in schools, even though the structure of the school was organized hierarchically. Social network analyses are suited to identify for individual schools via which and via how many others persons are connected in a network. These relationship structures represent the potential to spread content. In this regard, communication and decision paths as well as cooperation and power structures, for example, can be analysed as microprocesses with social network analysis. In addition to indirect effects, social network analysis can also be used to identify reciprocal effects, and in which schools teachers are connected only unidirectionally (person A chooses person B, but person B does not choose person A) or mutually and thus reciprocally (person A chooses person B, and person B chooses person A).

Indirect relations were also identified by Maag Merki et al. (Chap. 12). Multilevel analysis of the log data revealed that the relation between teachers' ratings of the day's activities and their daily satisfaction varied school-specifically and that it was moderated by teachers' interests in assessment and further development of their own teaching practices. Although these findings need to be tested in larger samples, they show the potential of log data to reveal differential and indirect effects. Complementary qualitative analyses could provide greater depth, such as was done in the study by Vanblaere and Devos (Chap. 10). In this way, explanations can be found that help to further develop theoretical models.

13.4 Variety of Meaningful Factors

To understand and assess school improvement processes, it is important to take a broad view of possible dimensions, structures, processes, and effects. Nevertheless, current school improvement research has strongly built on well-established dimensions and empirical findings (such as leadership practices or cooperation), which resulted in limited variability in research focus, and this has possibly limited development of more fully understanding the mechanisms involved in school improvement. An interesting extension of research on school leadership is presented in the contribution by Lowenhaupt (Chap. 8). In the study, the focus is on a linguistics method for analysing the rhetoric of school leaders. Lowenhaupt discovered that the rhetoric that school leaders use varies, and that rational, ethical or affective aspects are emphasized depending on the situation. As such, school leaders aim to initiate or influence development processes and school practices by differentiating their rhetoric. It would be interesting to investigate how differing rhetorical means affect teachers' motivation or interest in reflecting on their own practice in terms of quality development, how rhetorical means covary with individual characteristics, or how their availability and use change over time. The methodology can be linked to neo-institutional theories (DiMaggio & Powell, 1983/1991) or micropolitical theories for assessment of organizations (Altrichter & Moosbrugger, 2015). As such, it

allows differentiated analysis of power structures, negotiation processes on goals, values, and norms, and it can provide a better understanding of why school reforms do not, or only partially, achieve desired aims. In this sense the methodology presented holds potential for future school improvement research and for studies assessing intended and unintended effects of accountability approaches.

13.5 Concluding Remarks

The illustrative studies in this volume show how innovative methodologies can enrich school improvement research and help further development thereof. Taken together, they also provide an overview that can be used to systematically select the kind of methodology that fits a certain aspect of school improvement best. Moreover, we think that multimethod designs in which the presented methodologies are combined with other, especially qualitative, methodologies are very promising to better understand the complex interplay between actors' subjective meanings, their attributions, motivations, and orientations (e.g. Weick, 1995), individual and collective actions, and school structures and educational systems.

The methodologies presented in this volume for studying school improvement processes in the context of complex education systems cannot claim to revolutionize school improvement research, especially because the contributions could only selectively address previous research gaps. In addition, investigation of, for instance, differential paths and nonlinear trajectories could not be included. Still, we hope that with the presented innovative methodologies and designs, as well as the resulting new perspectives, we have provided inspiration for the study of school improvement as a multilevel, complex, and dynamic phenomenon. Future studies on key aspects thereof will provide a deeper understanding of school improvement in the context of societal and professional demands, and this will have a positive effect on the quality of school organisation, instruction, and ultimately on student learning.

References

- Altrichter, H., & Moosbrugger, R. (2015). Micropolitics of schools. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (Vol. 21, 2nd ed., pp. 134–140). Oxford, UK: Elsevier.
- Anusic, I., Lucas, R. E., & Donnellan, M. B. (2016). The validity of the day reconstruction method in the German socio-economic panel study. *Social Indicators Research*, 1–20. <https://doi.org/10.1007/s11205-015-1172-6>
- Decuyper, S., Dochy, F., & Van den Bossche, P. (2010). Grasping the dynamic complexity of team learning: An integrative model for effective team learning in organisations. *Educational Research Review*, 5(2), 111–133.
- DiMaggio, P. J., & Powell, W. W. (1983/1991). The Iron cage revisited: Institutional isomorphism and collective rationality. In W. W. Powell & P. J. DiMaggio (Eds.), *The new institutionalism in organizational analysis* (pp. 63–82). Chicago, IL: University Chicago Press.

- Eggen, T. J. H. M. (2008). Adaptive testing and item banking. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 215–234). Göttingen, Germany: Hogrefe.
- Feldhoff, T., Radisch, F., & Klieme, E. (2014). Methods in longitudinal school improvement research: State of the art. *Journal of Educational Administration*, 52(5).
- Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Oakland, CA: University of California Press.
- Gräsel, C., Fußangel, K., & Pröbstel, C. (2006). Lehrkräfte zur Kooperation anregen – eine Aufgabe für Sisyphos? *Zeitschrift für Pädagogik*, 52(6), 205–219.
- Hallinger, P., & Heck, R. H. (1998). Exploring the principals' contribution to school effectiveness: 1980-1995. *School Effectiveness and School Improvement*, 9(2), 157–191.
- Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, 20(1), 102–116.
- Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques* (Quantitative methodology series, 2nd ed.). New York, NY: Routledge.
- Kyriakides, L. (2007). Generic and differentiated models of educational effectiveness. In T. Townsend (Ed.), *International handbook on school effectiveness and improvement* (pp. 41–56). Dordrecht, The Netherlands: Springer.
- Luyten, H., & Sammons, P. (2010). Multilevel modelling. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 246–276). Abingdon, UK/New York, NY: Routledge.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60, 577–605.
- Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23(3), 187–194.
- Moolenaar, N. M., Slegers, P. J. C., & Daly, A. J. (2012). Teaming up: Linking collaboration networks, collective efficacy, and student achievement. *Teaching and Teacher Education*, 28(2), 251–262.
- Ohly, S., Sonnentag, S., Niessen, C., & Zapf, D. (2010). Diary studies in organizational research. An introduction and some practical recommendations. *Journal of Personnel Psychology*, 9(2), 79–93. <https://doi.org/10.1027/1866-5888/a000009>
- Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222). New York, NY: Cambridge University Press.
- Spillane, J. P., Hopkins, M., & Sweet, T. M. (2015). Intra- and interschool interactions about instruction: Exploring the conditions for social capital development. *American Journal of Education*, 122(1), 71–110.
- Spillane, J. P., Shirrell, M., & Adhikari, S. (2018). Constructing “experts” among peers: Educational infrastructure, test data, and teachers' interactions about teaching. *Educational Evaluation and Policy Analysis*, 40(4), 586–612.
- Stoll, L., & Fink, D. (1996). *Changing our schools: Linking school effectiveness and school improvement*. Buckingham, UK: Open University Press.
- Valsiner, J., Molenaar, P. C., Lyra, M. C., & Chaudhary, N. (Eds.). (2009). *Dynamic process methodology in the social and developmental sciences*. New York, NY: Springer.
- Van den Noortgate, W., Opendakker, M. C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, 16(3), 281–303.
- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 21, 1–19.
- Weick, K. E. (1995). *Sensemaking in organizations*. London, UK: Sage.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

