Myint Swe Khine *Editor*

Methodology for Multilevel Modeling in Educational Research

Concepts and Applications

Springer

Methodology for Multilevel Modeling in Educational Research

Myint Swe Khine
Editor

Methodology for Multilevel Modeling in Educational Research

Concepts and Applications

Springer

Editor
Myint Swe Khine
Curtin University
Bentley, WA, Australia

ISBN 978-981-16-9141-6 ISBN 978-981-16-9142-3 (eBook)
https://doi.org/10.1007/978-981-16-9142-3

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 Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Part I Introduction

1 Hierarchical Linear Modeling and Multilevel Modeling in Educational Research 3
Myint Swe Khine

Part II Theoretical Foundations and Conceptual Frameworks

2 A Primer for Using Multilevel Confirmatory Factor Analysis Models in Educational Research 11
Christine DiStefano and Tiejun Zhang

3 Multilevel Model Selection: Balancing Model Fit and Adequacy 29
D. Betsy McCoach, Sarah D. Newton, and Anthony J. Gambino

4 Concepts and Applications of Multivariate Multilevel (MVML) Analysis and Multilevel Structural Equation Modeling (MLSEM) 49
Yang Yang, Mengchen Su, and Ren Liu

5 Data Visualization for Pattern Seeking in Multilevel Modeling 69
Chong Ho Alex Yu

6 Doubly Latent Multilevel Structural Equation Modeling: An Overview of Main Concepts and Empirical Illustration 99
Irena Burić

Part III Methodology for Multilevel Modeling

7 Analyzing Large-Scale Assessment Data with Multilevel Analyses: Demonstration Using the Programme for International Student Assessment (PISA) 2018 Data 121
Julie Lorah

8 Multilevel Modelling of International Large-Scale Assessment Data 141
Anastasios Karakolidis, Vasiliki Pitsia, and Jude Cosgrove

9 Transparency and Replicability of Multilevel Modeling Applications: A Guideline for Improved Reporting Practices 161
Wen Luo, Eunkyeng Baek, and Haoran Li

10 Application of Multilevel Models to International Large-Scale Student Assessment Data 185
Maciej Jakubowski and Tomasz Gajderowicz

Part IV Multilevel Analysis of PISA and TIMSS Data

11 Changing Trends in the Role of South African Math Teachers' Qualification for Student Achievement: Findings from TIMSS 2003, 2011, 2015 205
Heike Wendt, Daniel Kasper, and Caroline Long

12 Revisiting the Relationship Between Science Teaching Practice and Scientific Literacy: Multi-level Analysis Using PISA 235
Hyesun You

13 Family Meals and Academic Performance: A Multilevel Analysis for Spain 255
Nerea Gómez-Fernández and Juan-Francisco Albert

14 Multilevel Modeling of Nordic Students' Mathematics Achievements in TIMSS 2019 271
Marie Wiberg

15 Teachers' Perceptions of School Ethical Culture: The Implicit Meaning of TIMSS 283
Orly Shapira-Lishchinsky

Part V Multilevel Modeling in Educational Research

16 Why They Want to Leave? A Three-Level Hierarchical Linear Modeling Analysis of Teacher Turnover Intention 311
Lixia Qin

17 Daycare Centers' Composition and Non-native Children's Language Skills at School Entry: Exploring the Nature of Context Effects Using Multilevel Modeling 339
Nina Hogrebe and Anna Marina Schmidt

18 Gender Effect at the Beginning of Higher Education Careers in STEM Studies: Does Female Recover Better Than Male? 363
Antonella D'Agostino, Giulio Ghellini, and Gabriele Lombardi

19 Service Satisfaction and Service Quality: A Longitudinal and Multilevel Study of User Satisfaction with Kindergartens in Norway . . . 383
Håvard Thorsen Rydland, Åsta Dyrnes Nordø, and Dag Arne Christensen

20 Multilevel Modeling and Assessment of the Study-Relevant Knowledge of First-Year Students in a Master's Program in Business and Economics . . . 403
Susanne Schmidt, Olga Zlatkin-Troitschanskaia, and Marie-Theres Nagel

Index . . . 425

Editor and Contributors

About the Editor

Dr. Myint Swe Khine, Ph.D., Ed.D. currently teaches at the School of Education, Curtin University, Australia. Prior to this appointment, he worked at the National Institute of Education, Nanyang Technological University, Singapore, and was a Professor and Chair of the Assessment and Evaluation Centre at the Emirates College for Advanced Education in the United Arab Emirates. His research interests are teacher education, learning sciences, educational measurement and assessment, and quantitative research. He has published widely in international refereed journals and edited several books. One of his recent books, *Academic Self-efficacy: Nature, Assessment, and Research*, was published by Springer in 2022.

Contributors

Juan-Francisco Albert Universitat de València, Valencia, Spain

Eunkyeng Baek Texas A&M University, College Station, Texas, USA

Irena Burić University of Zadar, Zadar, Croatia

Dag Arne Christensen NORCE Social Science - Norwegian Research Centre, Bergen, Norway

Jude Cosgrove Educational Research Centre, Dublin, Ireland

Antonella D'Agostino University of Napoli Parthenope, Napoli, Italy

Christine DiStefano University of South Carolina, Columbia, SC, USA

Tomasz Gajderowicz University of Warsaw, Warszawa, Poland

Anthony J. Gambino Department of Educational Psychology, University of Connecticut, Storrs, CT, USA

Giulio Ghellini University of Siena, Siena, Italy

Nerea Gómez-Fernández Universitat Politècnica de València, Valencia, Spain; Centro Universitario EDEM-Escuela de Empresarios, Valencia, Spain

Nina Hogrebe Hamburg University of Applied Sciences, Hamburg, Germany

Maciej Jakubowski University of Warsaw, Warszawa, Poland

Anastasios Karakolidis Educational Research Centre, Dublin, Ireland

Daniel Kasper University of Hamburg, Hamburg, Germany

Myint Swe Khine Curtin University, Perth, WA, Australia

Haoran Li Texas A&M University, College Station, Texas, USA

Ren Liu University at Buffalo, Buffalo, NY, USA

Gabriele Lombardi University of Siena, Siena, Italy

Caroline Long University of Johannesburg, Gauteng, South Africa

Julie Lorah Indiana University, Bloomington, IN, USA

Wen Luo Texas A&M University, College Station, Texas, USA

D. Betsy McCoach Department of Educational Psychology, University of Connecticut, Storrs, CT, USA

Marie-Theres Nagel Johannes Gutenberg University, Mainz, Germany

Sarah D. Newton Department of Educational Psychology, University of Connecticut, Storrs, CT, USA

Åsta Dyrnes Nordø NORCE Social Science - Norwegian Research Centre, Bergen, Norway

Vasiliki Pitsia Educational Research Centre, Dublin, Ireland

Lixia Qin University of Wisconsin System, Madison, WI, USA

Håvard Thorsen Rydland NORCE Social Science - Norwegian Research Centre, Bergen, Norway

Anna Marina Schmidt University of Münster, Münster, Germany

Susanne Schmidt Johannes Gutenberg University, Mainz, Germany

Orly Shapira-Lishchinsky Bar-Ilan University, Ramat Gan, Israel

Mengchen Su University at Buffalo, Buffalo, NY, USA

Heike Wendt University of Graz, Graz, Austria

Marie Wiberg Department of Statistics, USBE, Umeå University, Umeå, Sweden

Yang Yang Beijing Normal University, Zhuhai, China

Hyesun You University of Texas, Austin, TX, USA

Chong Ho Alex Yu Azusa Pacific University, Azusa, CA, USA

Tiejun Zhang University of South Carolina, Columbia, SC, USA

Olga Zlatkin-Troitschanskaia Johannes Gutenberg University, Mainz, Germany

Part I
Introduction

Chapter 1
Hierarchical Linear Modeling and Multilevel Modeling in Educational Research

Myint Swe Khine

Abstract Hierarchical linear modeling, also known as multilevel modeling, is increasingly prevalent in social science research because of its advantages not available in traditional statistical analysis. Education research often involves using data of nested nature since schooling systems are in a hierarchical structure. The students are clustered in the classrooms, classrooms are clustered in the schools, and schools are clustered in districts. Statistical techniques that account for the hierarchy are more suitable and accurate for analyzing such hierarchical data. Multilevel modeling provides a range of possibilities in analyzing complex data. Substantial numbers of studies have shown the flexibility of multilevel analysis and reported the unique advantages in examining intricate educational issues. This chapter synthesizes the studies reported in this book and describes the applicability of multilevel modeling in educational research.

Keywords Hierarchical linear modeling · Multilevel modeling · International large-scale studies · PISA · TIMSS · R package

1.1 Introduction

Educational research frequently involves problems investigating the relationships between students, classrooms, and school contexts. Such schooling systems present an example of hierarchical structure, with students clustered within classrooms, which themselves are clustered within the schools. The data gathered from the educational settings are hierarchical in nature, and all the observations are nested within the individuals at multiple levels. Analysis of hierarchical data is best performed using statistical techniques that account for the hierarchy. Multilevel modeling affords a range of possibilities for asking questions of the data that cannot be adequately addressed using traditional analytical methods. A substantial body of literature has

M. S. Khine (✉)
Curtin University, Perth, WA, Australia
e-mail: m.khine@curtin.edu.au

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_1

shown the versatility of multilevel analysis and elucidated the unique advantages in examining complex and wide-ranging educational issues.

Many investigations have been conducted and disseminated in the literature, and studies related to multilevel modeling of educational data are becoming prevalent. This volume aims to document recent attempts to conduct systematic and prodigious research using multilevel analysis in educational settings, and share their findings and identify future research directions. The book brings together leading experts around the world to share their outstanding and exemplary works in the field, detailing the recent advances, creative and unique approaches, and innovative methods using multilevel modeling and theoretical and practical aspects of multilevel analysis in culturally and linguistically diverse educational contexts.

This book is organized into four parts. The first part of the book presents theoretical foundations and conceptual frameworks related to multilevel modeling. The second and third parts of the book cover methodology for multilevel modeling and multilevel analysis of Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) data. In the final part of the book, educational research with large-scale data using multilevel analysis is presented.

1.2 Theoretical Foundations and Conceptual Frameworks

Five chapters in Part I cover theoretical foundations and conceptual frameworks for multilevel modeling. In Chap. 2, DiStefano and Zhang from the University of South Carolina demonstrate using multilevel confirmatory factor analysis in educational research. The authors express that confirmatory factor analysis is a technique to test a measurement framework to provide evidence of construct validity. The chapter discusses the multilevel confirmatory factor analysis (MCFA) framework and offers assistance for researchers interested in using this framework. The advantages of accommodating the multilevel framework and the effects of MCFA on parameter estimates and fit indices are detailed.

In Chap. 3, McCoach, Newton, and Gambino observed that multilevel model selection is critical in statistical modeling. The authors review the multilevel modeling estimation techniques and present model selection criteria and framework for decomposing outcome variance to measure model adequacy. The chapter also provides recommendations for the multilevel modeling selection process. Yang, Su, and Liu introduce concepts and applications of the state-of-the-art procedures in multilevel modeling in Chap. 4. The authors describe the current research trends that consider systematical thinking of the relations in education and continuous measurement of students' performance. The chapter offers two techniques, Multivariate Multilevel (MVML) Analysis and Multilevel Structural Equation Modeling (MLSEM). While the MVML approach allows for multiple outcomes analysis simultaneously, MLSEM offers more appropriate estimates by considering the intra-class correlations. The authors discuss with examples how these models can be used in educational research.

Alex Yu's chapter (Chap. 5) explains the visualization in analyzing international large-scale assessment data. The author suggests that in such large-scale data analysis of data from Programme for International Student Assessment (PISA), Trends for International Math and Science Study (TIMSS), and High School and Beyond (HSB), referring to *p*-value is not sufficient, and data visualization and pattern seeking are recommended. This chapter demonstrates how various data visualization techniques can extract insight from data at each step of multilevel modeling. In particular, the chapter offers procedures such as linking and brushing, binning and median smoothing, usage of a bubble plot, local filter, analysis of a mean plot, residual plot, and other related methods. In Chap. 6, Burić introduces doubly latent multilevel structural equation modeling (DL-MSEM). The author asserts that this approach enables testing theoretically relevant relationships at a proper level of analysis at class, teacher, and school levels and controlling for measurement by using multiple indicators for latent variables at student and teacher levels. The approach also addresses the sampling error by incorporating the scores for different students in the same class as multiple indicators of latent variables at the teacher level. The author examines the associations between teachers' self-efficacy, quality of instruction, and their students' motivational beliefs using the DL-MSEM approach. The chapter details the steps involved in the analysis and interpretation of the results.

1.3 Methodology for Multilevel Modeling

Part II contains four chapters and covers methods for multilevel modeling to analyze large-scale data from different perspectives. In Chap. 7, Lorah demonstrates a step-by-step procedure for analyzing large-scale data such as Programme for International Student Assessment (PISA) with multilevel modeling. The author discusses the complexity of survey design typically used in large-scale data gathering, involving sampling and replicating weights and plausible values. The chapter offers example analysis of PISA data, model specification, and analysis using the R package. The interpretations of the results and annotated output would help the readers understand the processes involved in multilevel modeling of PISA data.

Karakolidis, Pitsia, and Cosgrove from the Educational Research Centre in Dublin (Chap. 8) offer an introduction and starting point for applying multilevel modeling of international large-scale data for education research and policy decisions from different perspectives. The authors believed that multilevel models are suitable for longitudinal and cross-sectional analysis for national and international large-scale studies due to the hierarchical and clustered nature of the data. The chapter explains the configuration of multilevel models, sampling weights, and plausible values typically found in the dataset. The authors also mention the different types of software available for multilevel analysis, including HLM, MLwiN, Mplus, SAS, and R.

From Texas A&M University, Luo, Baek, and Li discuss the transparency and replicability of multilevel modeling applications and provide a guideline for improved reporting practices in Chap. 9. The authors have an opinion that reporting practices in multilevel modeling applications in education and psychology lack clarity and completeness in areas such as reliability and validity of multilevel measures, model specifications, description of missing data mechanisms, power analyses, assumption checking, model comparisons, and effect sizes. Using real-life research data examples, the authors illustrate the reporting principles and guidelines and a checklist in their chapter. From another viewpoint in Chap. 10, Jakubowski and Gajderowicz from the University of Warsaw highlight the importance of plausible values, survey weights, and replicate weights in analyzing the international large-scale student assessment. They also discuss the applications of complex models such as three-level models and models with cross-level effects. They explored the relationship between socioeconomic background and student achievement in PISA 2018 to demonstrate the analysis and interpret the results.

1.4 Multilevel Analysis of PISA and TIMSS Data

In Part III of the book, the researchers from Austria, South Africa, Germany, the USA, Spain, Sweden, and Israel share their findings from various research projects on PISA and TIMSS data analyses. In Chap. 11, Wendt, Kasper, and Long describe the changing trends in the role of South African mathematics teachers' qualification enhancement for student achievement. The study used TIMSS 2003, 2011, and 2015 datasets to conduct multilevel regression analysis. The results point that South Africa has made substantial progress in uplifting teachers' formal qualification levels in their education system. It was also found that the teacher's formal level of education is in general not significantly associated with students' mathematics achievement.

Hye Sun You shares the results from her study to examine the relationship between science teaching practice and scientific literacy in Chap. 12. The author draws the data from PISA 2015 and conducted the multivariate multilevel analysis of 5712 American students from 177 schools and two teaching practices: inquiry-based teaching and direct instruction. After controlling for student- and school-level variables, the multilevel modeling results showed that inquiry-based teaching was significantly negatively related to scientific literacy. It was also found that direct instruction was significantly positively associated with scientific literacy. The author concluded that findings from the study would help understand the in-depth knowledge about science teaching practices and student performance. In Chap. 13**,** Gómez-Fernández and Albert from Spain write about the relationship between family meals and academic performance. The authors are interested to understand whether the frequency of shared family meals has any connection with the academic performance of adolescence. Using PISA 2015 data, the authors conducted a multilevel analysis. The results revealed a positive relationship between the frequency of parents eating the main meal with their children and academic performance in reading comprehension.

Marie Wiberg from Umeå University in Sweden (Chap. 14) investigated the Nordic students' mathematics achievements in TIMSS 2019, which involves grade 8 students in Finland, Norway, and Sweden. The research questions include which school-level factors are associated with mathematics achievement in Nordic countries and whether the identified school-level factors are the same or differ according to low or high effective schools. The research also examined whether the identified school-level factors are country-specific or similar between countries. Considering that different countries have different education systems, it was found that various aspects apply to different countries. In another study, Shapira-Lishchinsky (Chap. 15) describes findings from the study to explore the teachers' perception of school ethical culture using the TIMSS 2015 database. The focus of the investigation is whether there is a shared perception of School Ethical Culture (SEC) among 45 participating countries. The author highlights that the TIMSS teachers questionnaire has additional meaning by identifying four dimensions. These include teachers' profession, care for students' learning, interaction with colleagues, and respect for rules. The author concluded that the impact of SEC dimensions would depend on each country's context and specific situations.

1.5 Multilevel Modeling in Educational Research

The last part (Part IV) of the book contains five chapters that address the multilevel modeling applications in educational research. The part begins with Chap. 16, in which Qin viewed that researchers are developing a more complex understanding of social phenomena and interactions using multilevel lenses. The chapter presents the joint impact and the interactive effects of individual and situational factors on teacher turnover and the variation across 32 countries considering teacher-, school- and country-level factors. The author demonstrated how teaching force research could significantly benefit from the multilevel analysis that affords to investigate teachers' macro/micro effects more explicitly. In Chap. 17, Hogrebe and Schmidt explore the nature of context effects in the composition of daycare centers and con-native children's language skills at school entry using multilevel modeling. The author notes that the composition of daycare centers, such as the proposition of children living in poverty, affects children's competencies, although the exact relationship is not clear.

In Chap. 18, D'Agostino, Ghellini, and Lombardi attempted to evaluate the gender effect of higher education careers in Science, Technology, Engineering, and Mathematics (STEM) studies using multilevel modeling. The study examines the 'transfer shock' which causes a temporary decrease in academic performance when students are transferred from one institution to another. The study reports the effect of transfer shock in this specific point of the students' career from a gender perspective. From the Norwegian Research Centre, NORCE, Rydland, Nordø, and Christensen (Chap. 19) report the longitudinal and multilevel study of the user satisfaction with kindergartens in Norway. The authors noted that user satisfaction surveys are typical in public administration as a quality measure and governance tool, including kindergarten education. By using the data from the Norwegian kindergarten survey (2016–2019), their study explores whether quality measures (staffing and staff's education) impact satisfaction and whether specific users (parents with the youngest children) are sensitive to service quality changes.

Finally, In Chap. 20, Schmidt, Zlatkin-Troitschanskaia, and Nagel from Germany present their findings on the assessment of study-relevant knowledge of first-year students in a master's degree program in business and economics. The study used the data collected in the third survey in the WiwiKom project, involving 1,523 master's students of economics at 27 universities and 13 universities of applied sciences. The authors explore whether the test for assessing business and economics knowledge used in studies focusing on bachelor's programs allows for the valid assessment and measurement of the domain-specific knowledge of students at the beginning of their master's studies. The authors note that the multilevel approach and latent analysis are the most suitable methods to explain the research questions in this context.

1.6 Conclusion

This introductory chapter provides an overview of the chapters included in the book. The chapters in this book portray conceptual, methodological, and analytical techniques in using multilevel models in educational research. Individually, each chapter extends our knowledge about specific aspects of multilevel modeling and its application to the educational context. The chapters in this book written by the practitioners enhance our understanding of multilevel modeling, recent development, and applications. The authors presented the state-of-the-art and novel approaches in data analysis in an easily accessible way. It is hoped that readers will benefit from the insightful accounts of understanding the concept, analytical tools, and interpretation of multilevel modeling and applications in educational research.

Dr. Myint Swe Khine, Ph.D., Ed.D. currently teaches at the School of Education, Curtin University, Australia. Prior to this appointment, he worked at the National Institute of Education, Nanyang Technological University, Singapore, and was a Professor and Chair of the Assessment and Evaluation Centre at the Emirates College for Advanced Education in the United Arab Emirates. His research interests are teacher education, learning sciences, educational measurement and assessment, and quantitative research. He has published widely in international refereed journals and edited several books. One of his recent books, *Academic Self-efficacy: Nature, Assessment, and Research*, was published by Springer in 2022.

Part II
Theoretical Foundations and Conceptual Frameworks

Chapter 2
A Primer for Using Multilevel Confirmatory Factor Analysis Models in Educational Research

Christine DiStefano and Tiejun Zhang

Abstract Confirmatory factor analysis is a popular analysis technique in educational research, often used to test a measurement framework or to provide evidence of construct validity. However, when nested data are present, dependencies in the data should be taken into consideration in order to produce accurate results. In this chapter, we discuss the multilevel confirmatory analysis (MCFA) framework, ways to accommodate multilevel data with latent variable estimation, and the effects of MCFA on parameter estimates and fit indices. An applied example is included. We conclude with recommendations for extending into the multilevel structural equation modeling framework and suggestions for future research in this area.

Keywords Multilevel Confirmatory Factor Analysis · Nested data · Design-based approach · Model-based procedure · Design effects

2.1 Introduction

Confirmatory factor analysis (CFA) is an important analysis tool, widely used in educational research as well as many other social science disciplines. The method is part of the structural equation modeling framework, where a hypothesized model is constructed a priori based on theory or prior evidence, and the fit of a model to a set of data from the population is evaluated through examination of a variety of model-data fit indices, parameter estimates, and other supports, such as model modification estimates, residual values, and standard errors of parameter estimates (e.g., Brown, 2015; Kline, 2016). As an analysis strategy, CFA is routinely employed to provide support for construct validity, scale refinement, tests of measurement invariance, and to evaluate alternative conceptualizations of a theoretical framework (e.g., Bandalos, 2018; Benson, 1998; Kline, 2016). The *Standards for Psychological and Educational Testing* recommend that the structure of an instrument should align with its scoring

C. DiStefano (✉) · T. Zhang
University of South Carolina, Columbia, SC, USA
e-mail: distefan@mailbox.sc.edu

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_2

procedures. Thus, CFAs may also be used to support a recommended a test's scoring algorithm (APA, AERA, NCME, 2014).

While CFA is useful, researchers also need to consider the data used for analyses. Nested or clustered data are a common occurrence with many educational research scenarios. For example, students may be nested in classrooms and schools nested within districts. Such situations pose complications because data collected from the same cluster are considered dependent. In other words, children within the same classroom share similar characteristics, such as similarities in instructional activities resulting from being taught by the same teacher. Even though other multilevel modeling procedures are popular in educational research, such as a regression framework (e.g., Hierarchical Linear Modeling) or an analysis of variance (ANOVA) framework (e.g., repeated measures designs) we focus on the use of CFA with multilevel data (MCFA) in educational research situations.

To gain greater understanding of the extent to which CFA and MCFA are currently used with empirical studies, a small-scale review was conducted for the past 10-years. Using the PsychInfo and Education Resources Information Center Database (ERIC) with the delimiter "confirmatory factor analysis", over 18,000 peer-reviewed academic journal articles CFA since 2011.[1] Given the frequency with which CFA has been employed in education research and other social science disciplines, this analysis tool and its benefits are well-known to researchers.

However, using the same databases and searching with delimiters of "multilevel" AND "confirmatory factor analysis," the number of peer-reviewed articles published in the past 10 years resulted in only 260 publications. This suggests that instead of accommodating the dependency among the data, the multilevel structure is typically ignored when conducting CFA. Reasons for ignoring nesting when using factor analysis may be due to: (a) limited availability of software packages that can automatically estimate a multilevel factor analytic structure, (b) convergence or estimation problems due to few clusters, or (c) failure to recognize the hierarchical nature of the data (e.g., Huang, 2017; Konold et al., 2014).

However, if nesting is present, it is important to accommodate the dependencies which occur in the data. Ignoring nesting can increase model misfit, produce biased parameter estimates, and attenuate standard errors of parameter estimates (e.g., Hox, 2010; Julian, 2001; O'Connell & McCoach, 2008; Pornprasertmanit et al., 2014). Multilevel analyses can allow researchers to investigate complex research questions which are relevant to different levels of the design structure (Huang & Cornell, 2016; Stapleton et al., 2016). Further, advances with software packages allow MCFA to be more easily accessed through many of the same programs that conduct confirmatory factor analyses (e.g., MPlus, R, LISREL, EQS).

In this chapter, we describe MCFA and the effects of ignoring nesting, conceptual principles related to handling nesting, and issues related to fit and model interpretation. Our focus is on the use of cross-sectional data where individuals are nested within clusters. For our discussion, a two-level MCFA will be examined; however,

[1] PsychInfo and ERIC searchers were current for period of September 2011 through September 2021.

the information extends to models with more than two levels. Finally, we provide an applied example to illustrate steps in testing an MCFA design and suggestions for future research.

2.1.1 Conceptual Principles

As with other parametric tests, confirmatory factor analysis (CFA) requires assumptions to be met to ensure valid inferences can be made. Assumptions related to data screening (e.g., multivariate normality, absence of outliers, choice of estimator in line with the metric level of data, etc.) are required. Considering characteristics unique to analysis of clustered data, CFA typically assumes observations in the sample data are independently and randomly sampled from the population. Breaking these two assumptions down, independence of observations means that the data should not be interconnected in any way. Therefore, when constructing the research design, data for analyses should be obtained through random sampling. Using a random sample effectively minimizes the dependence among subjects as each observation has an equal opportunity to be selected from the population. When conducting a single-level CFA, independence of observations is usually assumed, therefore the variance of a variable is mainly accounted for by individual differences.

Clustered data violates a key assumption necessary for a single-level CFA . Simply stated, members of the same group are more alike than members of different groups. Consider a situation where the goal is for students to assess the behavioral characteristics of their school. This is defined as a two-level model, where students are nested in schools. To examine how students vary in ratings of school climate, researchers may use individual observations; this is termed the within-level (or, alternatively, level-1, micro-level, low-level, first-level). We can also examine how perceptions of climate vary between schools. School-level investigations are considered the between-level (or level-2, macro level, higher level, second level, etc.). While this example includes two levels, multilevel data in educational and behavioral science settings may have more between-levels if the clusters are further nested in even larger unit(s). Following the same school climate example where students (level-1/within-level) are nested in schools (level-2/between-level), we could extend the example to consider that schools are nested within a school district (level-3), and school districts are nested within a state (level-4), and so on. Even with multiple "higher" levels, there is one within-level (level-1) but, depending upon the situation, there may be up to *n* between-levels.

As the MCFA data are clustered, it is important to estimate the level of dependence by interpreting the intraclass correlation coefficient (ICC). The index can be interpreted as the proportion of variance in an observed variable found at the cluster level rather than at the individual level (Stapleton et al., 2016). Thus, the ICC indicates how strongly cases within the same cluster are interdependent (e.g., Kline, 2016); the item-level ICC is calculated as:

$$\text{ICC} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

Examining the formula shows that item-level ICC is the proportion of variability at the between level compared to the total variability (i.e., between- and within-level variability). Considering the components in the equation, the between-level variability is the cluster or group-level variance component, representing the deviation of group/cluster means from the grand mean across all groups. The within-level variance is a pooled estimate of average within-level deviation of scores within the groups after accounting for sample size; the total variance is the sum of the group-level and within-level variance components (Muthén, 1994). ICC values are computed at the item level but may be averaged across the set of items to provide an estimate of the amount of dependability in a set of items.

ICC estimates are ratios of variance; therefore, index values range from 0 to 1, with higher values indicating a greater proportion of between-level variance. In other words, larger ICC values suggest more dependency is present in the data. The values are often examined prior to analysis to determine if a multilevel analysis is warranted. The higher the value, the greater the amount of dependency in the data, and greater bias observed in model results if the multilevel nature of the data is not considered (Muthén, 1991). Normally, an ICC value greater than 0.05 suggests the need for a multilevel analysis (Julian, 2001; Pornprasertmanit et al., 2014; Stapleton, 2013). In practice, multilevel modeling may provide few benefits when ICCs are less than 0.05 (Muthén, 1994; Stapleton et al., 2016); however, research has suggested that even an ICC of 0.01 could have an impact with large clusters (Stapleton, 2013).

2.1.2 Benefits of Using the MCFA Framework

While many statistical techniques can accommodate nested data, there are advantages to using the CFA framework with multilevel data over other approaches. A primary benefit is that a CFA model may be constructed at each level and that models may be distinct at the different levels (Stapleton, 2013). These structures could have different interpretations and/or different factor structures depending on the level of analysis (Stapleton et al., 2016). This flexibility allows researchers to construct a model which best fits the nuances of the measurement at the different levels. In addition, the structural modeling framework can also handle multiple covariates and multiple outcome variables easily. Thus, models may be constructed to include multiple distal outcomes, where outcomes differ based on the level of the data. While other methods can also handle covariates and outcome variables, a benefit of the CFA framework is that multiple dependent variables may be included in one model, creating a parsimonious testing situation and, perhaps avoiding multiple testing errors.

2.2 Analysis Choices for MCFA

When constructing latent variable models, researchers have different choices for analyzing multilevel data. In this section, we discuss three options: ignoring nesting, using design effects and constructing a multilevel CFA using a pooled covariance matrix.

2.2.1 Ignoring Nesting

A single-level CFA can be applied to multilevel data, by essentially ignoring the multilevel structure and analyzing data at a given level. For example, the latent structure can be examined using disaggregated data (within-level) at the individual level, thereby ignoring higher level influences. Or data may be aggregated for analysis (e.g., average classroom scores) and the CFA estimated using summary data and ignoring individual responses. In both cases, a "single-level" CFA is tested. In the disaggregated approach, the between-level is ignored and affects the within-level variance. Therefore, the disaggregated variance of a variable in single-level CFA would be larger than its within-level variance in MCFA. Similarly, the variance of a variable using aggregated CFA will be greater than the group-level variance of the same variable in MCFA because the within-level variance is ignored and added to the aggregated variance.

When a MCFA is a more viable option, ignoring the multilevel nature may result in biased results—where the extent to which the findings are biased mainly depends on the magnitude of ICC and cluster size (i.e., the number of cases per cluster) (Julian, 2001; Moerbeek, 2004; Pornprasertmanit et al., 2014; Stockford, 2009). Overall, the disaggregated approach demonstrates attenuated model fit in the Chi-square statistic and root mean squared error of approximation (RMSEA) when the nested data structure is ignored, especially when ICC is large. The consequences of ignoring nesting are negligible when ICC values are low (e.g., 0.05) and the group size is small. Here, the Chi-square statistic is only minimally inflated, and the model parameters and their standard errors are essentially unbiased. The consequences are more severe when ICC values are large (e.g., above 0.5): the Chi-square global fit index is inflated, the unstandardized parameter estimates are overestimated, and the standard errors are underestimated (Julian, 2001; Pornprasertmanit et al., 2014). As unstandardized parameter estimates were found to always exhibit bias when ICC values are high, Pornprasertmanit et al. (2014) recommended that researchers focus on interpreting standardized estimates. If both the within- and between-levels of the CFA model exhibit similar parameter values, standardized parameter estimates are equal to the unstandardized parameters. However, if the parameter values differ substantially across levels, the standardized parameter estimates are likely to be biased as ICC values increase beyond 0.15. The standard errors of the standardized parameter estimates are either underestimated toward large macro-level communalities or overestimated under small micro-level communalities, but the bias is negligible when ICC is less than 0.05. In general, the disaggregated approach provides less biased estimates and standard errors of standardized estimates when ICC is 0.05 or lower, but the attenuated fit indices may result in rejection of acceptable models (Pornprasertmanit et al., 2014).

If data are aggregated, that is, the micro-level is ignored, the model fit indices (global Chi-square index, RMSEA) are the same as the partially saturated MCFA that allows all variables covary at the macro level. If the parameter values are the same at both the micro- and macro-level, the standardized estimates are not biased. However, if the parameter values differ across levels, the standardized estimates will show bias toward the micro-level standardized estimates, especially when ICC is lower than 0.25 or the group size is small. In terms of standard errors, Moerbeek (2004) indicated they are less biased, but Pornprasertmanit et al. (2014) argued that the bias in standard errors of standardized parameters is only ignorable when ICC is greater than 0.75. A possible explanation for this inconsistency may be due to the use of standardized parameters in Pornprasertmanit's study. The calculation of standardized parameters involves unstandardized parameters and indicator or factor variance. Aggregation would overestimate indicator or factor variance, which results in inaccurate standardized estimates. In general, the aggregated approach provides accurate parameters estimates and standard errors of standardized parameters when ICC is 0.75 or above, though this situation is not very likely in practice (Pornprasertmanit et al., 2014).

2.2.2 Accounting for Nested Data

Researchers recognizing the multilevel nature of the data generally follow one of two approaches: design-based and model-based procedures (Stapleton, 2013), when conducting an MCFA.

Design-Based Approach. A design-based approach recognizes that the data are nested in nature and that the dependency in the data needs to be accommodated into the design to produce accurate standard errors of parameter estimates and model fit (e.g., Julian, 2001; Stapleton, 2013). However, under this viewpoint, it is not of interest to answer questions that deal with more than one level, and typically, the focus of the analysis is at the within- or individual level. Under this approach, the nested structure of the data is considered part of the design process, or how data arise in practice. This process includes incorporating a "design effect" into the analyses to accommodate nesting.

To determine if the multilevel structure is adversely affecting CFA results, a design effect value may be computed. This value is a function of both the ICC and the average cluster size (Muthén & Muthén, 1998–2021). The design effect may be calculated as:

$$1 + (\text{average cluster size} - 1) \times \text{ICC},$$

As a rule of thumb, design effect values larger than 2.0 is thought as a benchmark, indicating that the clustering should be considered (Hox & Maas, 2002). However, using a design effect is not recommended if researchers have small cluster sizes (under 10) or are interested in examining relationships at the higher level(s) (Lai & Kwok, 2015).

To include a design effect, a CFA model is constructed "ignoring" the nested factor; however, instead of analyzing the model as a single-level CFA approach, a survey weight is included in the estimation process. Typically, the design factor helps by weighting the data according to the number of cases in a cluster at higher levels of the analysis. Such weights are commonly used with complex, large-scale surveys/databases to correct for the effects of unequal probabilities of case selection (Asparouhov, 2006). With MCFA, design effects may be incorporated and to treat clustered data as a nuisance. Under this situation, the standard errors are adjusted for by the sampling design. While this technique provides correct standard errors and properly accounts for dependence present in the data, it does not allow for an examination of between-cluster variance which is unaccounted for by predictors included in the model or models which may be of interest at higher levels (Asparouhov, 2006; Lai & Kwok, 2015; Stapleton, 2013).

Model-Based Procedures. Model-based procedures allow an examination of relationships between variables at the between- and the within-levels. In MCFA, the "usual" variance–covariance matrix input for analyses is separated into two parts: a pooled-within-group covariance matrix and a between-group covariance matrix. Thus, the variability is decomposed into different sources of variance aligned with the various levels of the MCFA. This process allows researchers to examine within-cluster and between-cluster relations simultaneously.

Model-based procedures allow researchers additional advantages for accommodating nested data including the flexibility to estimate and model between-cluster effects; increased power at the within-cluster level, and the possibility to examine the within-cluster relations among variables across clusters (Stapleton, 2013).

Figure 2.1 provides a diagram of an MCFA with two levels and six (X_1 - X_6) observed variables. The variance of each observed variable (e.g., X_1) is explicitly decomposed into a between-variance component (X_{1B}) and a within-variance component (X_{1W}). In Fig. 2.1, F1B and F2B are the theoretical latent variables that account for the between-level covariance among the observed variables. Similarly, F_{1W} and F_{2W} are the theoretical latent variables that drive the within-level covariance among the observed variables. In our example, the latent structure is the same across levels; that is, each level has the same two factors, and each factor was measured by the same variables.

With MCFA, the total population covariance matrix Σ_T is split into two parts—a within-covariance matrix, Σ_W, and a between-covariance matrix, Σ_B. Σ_W and Σ_B assess the within-cluster and between-cluster effects, respectively. These two matrices are orthogonal and additive. Due to these principles, the between-group relationships among variables do not have to be the same as the within-group relationship. Said another way, the CFA structure which is tested may differ at different levels of the design.

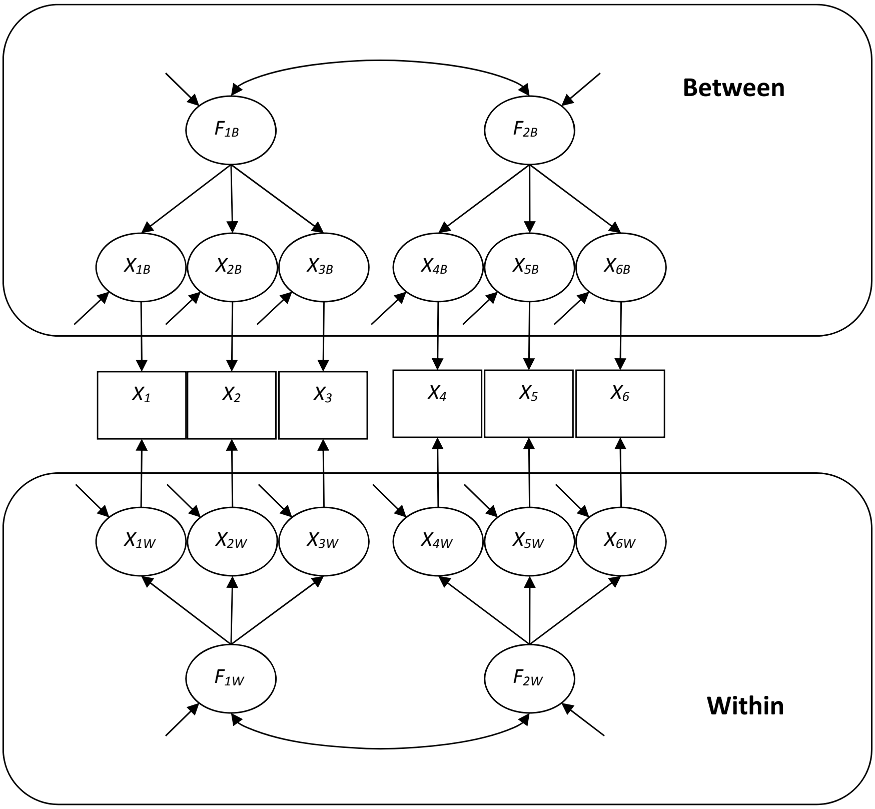


Fig. 2.1 Two-factor multilevel CFA model

The three population matrices used in MCFA are estimated by sample data. The sample total covariance matrix S_T is decomposed into S_W and S_B matrices, respectively. However, the pooled within-covariance matrix, $S_{PW,}$ is used to estimate Σ_W (instead of S_W) as the pooled matrix is an unbiased estimator of the population within-group matrix. The three sample covariance matrices S_T, S_B, and S_{PW} can be calculated by the following equations (see Muthén, 1993 for additional explanation).

$$S_T = (N-1)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{Ng} \left(y_{ig} - \overline{y}\right)\left(y_{ig} - \overline{y}\right)'$$

This equation shows that the sample covariance matrix can be calculated in the typical manner. S_T illustrates how an observed score y for individual i in group g deviates from the grand mean, $\overline{y}$

$$S_{PW} = (N - G)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{Ng} \left(y_{ig} - \overline{y}_g\right)\left(y_{ig} - \overline{y}_g\right)'$$

The pooled within-cluster matrix, $\mathrm{S}_{\mathrm{PW},}$, illustrates how an observed score *y* for individual *i* in group *g* deviates from the cluster group mean. The matrix is weighted by the total sample size (*N*) less the number of clusters.

$$S_B = (G - 1)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{Ng} \left(\overline{y}_g - \overline{y}\right)\left(\overline{y}_g - \overline{y}\right)'$$

Finally, the between-cluster matrix, S_B, illustrates how the mean of the cluster deviates from the grand mean. S_B is weighted by the cluster size.

A stepwise approach has been recommended to assess the necessity of multilevel procedures and evaluate relevant models; however, researchers have not reached an agreement on the specific steps to follow (e.g., Hox, 2002; Huang, 2017; Muthén, 1994; Stapleton, 2013). The stepwise processes generally begin with a single-level factor analysis of S_T followed by an estimation of the amount of between-level variation (ICCs) contained in the estimates. Additional steps differ depending on the procedures followed. These may include estimation of a model with parameters constrained to be equal across the level (i.e., null model; Hox, 2002; Huang, 2017); estimation of a saturated model, where all parameters are intercorrelated for a given level, with a model imposed at the other level(s) of a model (Stapleton, 2013); and/or estimation of random effect parameters at the within-level (Stapleton, 2013). Stepwise approaches to MCFA were often recommended due to difficulties encountered with employing multilevel procedures in software packages, estimation of ICC values, and the decomposition of Σ_T into different model levels.

Nowadays, the development of SEM software allows more complex multilevel models with Σ_{Wg} differing across clusters and it may be argued that "real world" practices for conducting MCFA do not follow a strict sequence of steps (Zyphur, 2019). Researchers may examine ICC levels and properties of variables prior to conducting a multilevel investigation to ensure that there is sufficient variability to model at the higher level(s). Further, researchers may reflect on the measurement to ensure that MCFA at group levels is theoretically and conceptually meaningful. Models may also be examined at individual and group levels separately before the hypothesized multilevel model is tested at both levels.

When using MCFA techniques, analysis problems may arise due to complexities of the situation. First, if there are not a sufficient number of clusters the macro levels of a design, the model may encounter estimation problems. In addition, most of the variability may be at the within-level and, due to the aggregation at the higher levels, more measurement error may be present (Zyphur, 2019). In such situations, item residual values may be very small due to high loading values at macro level(s) resulting in identification problems. If this is encountered, increasing the number of iterations required, fixing indicator variances to zero, and use of Bayes estimation techniques may offer a solution (Zyhpur, 2019).

2.2.3 Evaluating Model Fit in MCFA

As with other covariance modeling techniques, model-data fit evaluation is a vital part of analyses. Typically, researchers compare competing models (i.e., representations of the theory) and use various statistical indices and substantive knowledge to identify the optimal model (DiStefano, 2016). To arrive at the optimal model, both global and local fit of the tested MCFA may be examined. Global fit assesses the overall fit of the hypothesized model to the data, while local fit issues focus on examinations of the "finer grains" of a model, including parameter estimates, standard errors of parameter estimates, potential model modifications, and residual estimates (e.g., DiStefano, 2016; Kline, 2016). While both areas are important, acceptable global fit for a model is a necessary precursor to examination of local fit, as a poorly fitting model should not be interpreted.

Even though MCFA represents a multilevel structure, the fit indices and associated cutoff values used for model-data fit evaluation are the same as used with single-level CFA models. Thus, researchers generally rely on the commonly used fit indices along with the traditionally recommended cutoff values (e.g., Hu & Bentler, 1999). Popular fit indices used by education researchers include the overall Chi-squared test of exact fit, root mean squared error of approximation (RMSEA) and its accompanying confidence interval, standardized root mean square residual (SRMR), comparative fit index (CFI), and the Tucker-Lewis Index (TLI, or the Non-Normed Fit Index) (DiStefano & Hess, 2005; Jackson et al., 2009). Provided is a brief description of what each fit index measures in single-level sources for context when interpreting MCFA results; additional details may be obtained in other sources (e.g., DiStefano, 2016; Kline, 2016; Schumacker & Lomax, 2016).

The global Chi-square index tests the overarching hypothesis that the model exactly fits the data, where nonsignificant *p*-values illustrate acceptable model-data fit. However, Chi-square can be sensitive to characteristics such as large model size, large sample size. CFI and TLI compare the fit of the tested model to a baseline model with no structure; however, the indices differ in that the CFI helps to consider the number of paths tested (i.e., freed) in a model. For both indices, generally values at or above 0.95 indicate acceptable fit (Schumacker & Lomax, 2016). The root mean square error of approximation (RMSEA) recognizes that no empirical model will fit the data exactly, thus, values at or below 0.05 indicate a close-fitting model and values between 0.05 and 0.08 to indicate adequate fit. The standardized root mean square residual (SRMR) provides an estimate of the amount of error remaining in the variance–covariance matrix with values under 0.05 indicate acceptable fit. While these indices are the most commonly used with model evaluations, other fit indices, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) may provide useful information in the MCFA context to help choose between models when maximum likelihood estimation or unweighted least squares estimators (and robust variants) are used (e.g., Ene, 2020).

MCFA Model Fit and Ignoring the Multilevel Structure. While popular CFA fit indices are commonly evaluated, the interpretations may lead to inaccurate interpretations of a model because the multilevel structure is not considered (Hsu et al., 2015). If a CFA model is tested ignoring the multilevel nature of the data, the Chi-square test statistic for exact fit is provided across both levels. Ryu (2014) notes that the value of the default maximum likelihood fit function is an overall estimate; however, the value is weighted differently based on model misfit at level-1 (weighted by *N*–*J*, where *J* is the number of groups and *N* is the total sample size) and at level-2 (due to *J*). The amount of misfit in the Chi-square test statistic also affects the calculation of the CFI, TLI, and RMSEA, and the lack of fit observed at each level is not disentangled (Ryu, 2014). Also, as sample size is generally larger at the within-level, fit indices are dominated by the larger sample size and may not be sensitive to misspecifications at the between level (Hox, 2010; Ryu, 2014; Ryu & West, 2009). For example, single-level calculations of CFI and RMSEA were not able to identify model misspecification at the between-fit level (Ryu &West, 2009). Further, if fit index values are outside of the commonly used bounds, the poor fit does not aid researchers in understanding which level of the model (i.e., level-1, level-2, both, etc.) does not fit well.

MCFA Fit and Applying Design Effects. Another MCFA situation that uses single-level fit indices is when design effects are used. Applying design effects corrects standard errors of parameter estimates when the focus is to essentially “ignore” the multilevel structure and focus on the within effects (Lai & Kwok, 2015; Raykov & DiStefano, 2021). The effect of ignoring the design effect “rule” on standard errors of parameter estimates has been investigated using differing ICC levels and cluster sizes (Lai & Kwok, 2015); however, recommendations for evaluating single-level fit indices under MCFA with design effects have not been thoroughly examined.

MCFA Fit and Model-Based Procedures. When estimating an MCFA, researchers tend to use both the same model-data fit indices used with single-level CFA (e.g., Chi-square, CFI, TFI, RMSEA, SRMR) and apply the same cutoff criteria recommended for single-level analyses (Kim et al., 2016). However, when running multilevel analyses, the indices reported generally examine overall model fit instead of level-specific indices (Kim et al., 2016). Overall fit indices are influenced more heavily by the within-level as the sample size is higher at the within than at the between level(s). Thus, fit indices were not found to be sensitive to model misspecifications at the higher level(s) (Hsu et al., 2015; Ryu & West, 2009). In addition to sample size, ICC also has an effect on the model fit indices with the between level (Hsu et al., 2015) with TLI and RMSEA exhibiting more sensitivity to higher ICCs than CFI or SRMR.

While fit indices provide an overall assessment of model-data fit, the values do not appropriately evaluate the fit of an MCFA because the multilevel structure is not considered. As an MCFA contains both within- and between-level models, single-level CFA fit indices have a potential limitation in detecting the between-level model’s lack of fit. Thus, the typically used fit indices are more sensitive to misspecification in the within-level model (Hsu et al, 2015). In addition, if the fit indices indicate a well-fitting model, a researcher does not know if both the between-level and within-level model fit well or because if the indices are not able to detect the misspecification at the between-level (Kwon, 2011). Researchers need to inspect evaluation methods that can separately estimate each level of the MCFA to understand which part of the hypothesized model did not fit the data (Kim et al., 2016).

As a remedy, researchers should include level-specific fit indices (Kim et al., 2016). Under multiple group modeling, the Mplus program automatically computes an SRMR value for each group (Asparouhov & Muthén, 2018). Other fit indices are computed at the overall level.

If following an MCFA stepwise approach, a partially saturated (PS) model approach has been proposed to obtain level-specific fit indices (Ryu & West, 2009). This approach uses the saturated within-level or between-level model (i.e., the between-level model is a just identified model and the within-level model is tested as hypothesized model). A PS model can be obtained by correlating all the observed variables and allowing all the covariances or correlations to be freely estimated at the between-level or within-level model. The between-level Chi-squared value ($\chi^2_{\mathrm{PS\_B}}$) can be calculated by specifying a hypothesized between-level model and saturating the within-level model because the saturated within-level model Chi-square value will equal zero and will not contribute to the overall model Chi-square (Hsu et al., 2015). In this way, the $\chi^2_{\mathrm{PS\_B}}$ only reflects the model fit of the hypothesized between-level model (Hsu et al., 2015). After $\chi^2_{\mathrm{PS\_B}}$ is obtained, other between-level specific fit indices, such as $\mathrm{RMSEA}_{\mathrm{PS\_B}}$, $\mathrm{CFI}_{\mathrm{PS\_B}}$, and $\mathrm{TLI}_{\mathrm{PS\_B}}$, can be calculated because these fit indices are based on the Chi-square value. Similarly, a within-level specific Chi-square value ($\chi^2_{\mathrm{PS\_W}}$) and other within-level specific fit indices (e.g., $\mathrm{RMSEA}_{\mathrm{PS\_W}}$, $\mathrm{CFI}_{\mathrm{PS\_W}}$, and $\mathrm{TLI}_{\mathrm{PS\_W}}$, can also be computed. Thus, the different levels will be evaluated by different fit indices (Hsu et al, 2015; Ryu & West, 2009). Ryu and West's work (2009) has shown the effectiveness of these indices, where the within-level specific fit indices correctly indicated poor model fit at the within-level and between-level specific fit indices successfully detecting lack of fit in the between-group model. The partially saturated approach for computing fit indices, however, is not frequently employed with MCFA.

2.3 Applied Example

To illustrate MCFA in an empirical situation, we provide an applied example. The state of South Carolina administers annual school climate surveys in all public schools. Students, teachers, and parents are surveyed to assess schools' respective school climate performance. The results of the surveys are shared with principals and district administrators to gauge yearly progress. Further, climate information is provided on the annual school report cards and is also used to meet requirements of the state's accountability legislation.

Survey instruments are administered to all teachers and to students and their parents at the highest grade level at each school in the entire state (typically 5, 8, and 11), though students in Grade 12 are not targeted. The surveys include multiple Likert-scale items measured on a four-point scale with anchors of 1 = Disagree, 2 = Mostly Disagree, 3 = Mostly Agree, and 4 = Agree.

The multilevel structure of the student forms (Ene, 2020; Ene et al., 2018) and the teacher forms have been investigated in prior analyses (Ene et al., 2016, 2017a, 2017b). Our focus is on the parent climate survey at the high school level, as the parent responses have not yet been investigated in the MCFA framework.

The sample of parents analyzed completed the climate survey in the spring of the 2017–2018 school year. The dataset contained information from 230 public high schools and 5,255 parent responses from across South Carolina. The school climate survey collects limited demographic data in an effort to increase response rates and protect respondents' privacy. Concerning the population of students enrolled students during the 2017–2018 school year, 33.6% were black or African American, 50.6% were White, 9.7% were Hispanic or Latino, and over 6% identified as American Indian, Asian, Hawaiian, Pacific Islander, or biracial (South Carolina Department of Education, 2018). Female and male students composed 48.9% and 51.1% of the student population, respectively (South Carolina Department of Education, 2018). However, characteristics of the respondents to the Parent School Climate Survey (e.g., respondent age, relationship to high school student, race/ethnicity) are not collected.

Previous research on the South Carolina Parent Climate Survey has identified a four-factor structure. The factors identified include: (1) Learning Environment (5 items—e.g., satisfaction with resources and textbooks), (2) Social-Physical Environment (5 items—e.g., cleanliness of the school and school grounds), (3) School Safety (3 items—perceptions of safety on the school campus), and Home-School Relationships (8 items—e.g., communication/ sharing information between parents and teachers.) This four-factor structure has been replicated using exploratory factor analysis and confirmatory factor analysis over different datasets and different organizational levels. Previous research has used a two-level MCFA design to investigate teacher and student responses of school climate across various levels (elementary, middle, and high school grades) (Ene et al., 2016, 2017a, b 2018). Findings have identified a multi-factor solution at the within-level and a one-factor solution at the between-level.

Alternative models were investigated with high school parents' responses to the South Carolina Parent School Climate Survey. First, a single-level CFA was employed. This analysis ignored school effects to examine parents' perceptions of school climate, regardless of the high school that their child attended. The second model was also a single-level CFA; however, design effects were incorporated into the analysis to correct standard errors and parameter tests for dependencies in the data due to the clustered structure. This analysis essentially ignores school-level effects and "removes" them from consideration. Two different two-level MCFA designs were examined. The first design considered that the same four-factor structure underlying the parent survey was present at both the within- and the between-level. This conceptualization of climate would consider that the four factors were present for individuals as well as aggregated across high schools. The second MCFA tested a four-factor structure at the within-level and a one-factor structure at the between-level, stating that individuals recognized four distinct aspects of school climate, but across schools, only an overall evaluation of school climate was provided.

Mplus (version 8.7) was used for all CFA analyses with the robust ML estimator employing a Satorra-Bentler mean correction. Data were considered to be continuous as the level of univariate skewness and kurtosis were low (below |1.3| for all variables) and, as there were relatively few cases with missing data, mean imputation was conducted prior to analyses. To assess global fit, CFA fit information was reported for the overall fit of the model using the Chi-square fit index, CFI, TLI, RMSEA, and its associated 90% confidence interval and level-specific SRMR values were also reported with MCFA models. Local fit indices, including parameter estimates, standard errors, and modification indices were also examined.

An initial investigation of the data yielded ICC values ranging from 0.052 to 0.167 (average ICC of 0.101), indicating dependency in the parent responses for the high school clusters. The survey data also yielded a sufficient number of clusters at the school level, 230, with an average sample size of 22.85 parents per cluster. Using the average ICC and average sample size, the design effect was estimated to be 3.21. Based on the review of the data, its structure, and ICC values, a multilevel design is warranted.

Table 2.1 provides a summary of global model fit indices. Across the MCFA designs. The single-level, four-factor correlated CFA showed acceptable fit for fit

Table 2.1 Global fit indices for tested models

Model	Single-level CFA	CFA with design effects	MCFA 4-factor both levels	MCFA 4-factor within 1-factor between
Chi-square	2777.69	2566.29	4660.50	4088.05
df	183	183	387	372
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001
AIC	178,440	178,440	177,021	176,774
BIC	178,893	178,893	177,651	177,503
RMSEA	0.052	0.050	0.046	0.044
90% CI	(0.050 – 0.054)	(0.048 – 0.052)		
p-value	0.030	0.579		
CFI	0.954	0.957	0.934	0.943
TLI	0.947	0.950	0.929	0.936
SRMR/SRMR-W	0.036	0.036	0.036	0.039
SRMR-B			0.165	0.119

indices other than the global Chi-square. While incorporating design effects into the modeling process did not greatly improve most model fit indices, the RMSEA *p*-value, supported close-fit between the model and the data (Kline, 2016). Two MCFA models were tested. The two-level model with the four-factor structure imposed at the within- and between-levels initially yielded estimation problems. Inspection showed that the between-level residuals were low; item uniqueness values were constrained to zero and the number of iterations was increased (Zyphur, 2019). These methods helped the model achieve a solution. For comparison, a two-level MCFA was estimated, with a four-factor model at the within-level and a one-factor model at the between-level. Comparing across these two MCFA models, the SRMR-between was lower for the 1-factor between-level model (SRMR = 0.119) than for the 4-factor between-level model (SRMR = 0.165); however both estimates are outside of recommended bounds. SRMR within-level estimates were favorable for both MCFA models tested. Investigating across all four models tested, AIC and BIC values were lowest for the 1-factor between model.

Both the 1-factor between-level MCFA and the design effect models were thought to be acceptable representations of the Parent Climate Survey dataset; therefore, local fit was inspected for these models. The MCFA reported very high loading estimates for all values (0.69 to 0.91) and high intercorrelations between factors (0.61 to 0.85) at the within-level; however, it is noted that the levels are similar for the design effect model. Based on information from model fit, prior research, comparison of model fit, the MCFA with the design effect applied was selected. At the theoretical level, it recognizes that there are differing opinions for parents with students in different high schools across the state; this variability can be incorporated into the analyses to accommodate the dependency in the data.

2.3.1 Areas for Future Research

MCFA is becoming easier for applied researchers to use and, with increased use of the technique there are also many areas for future research. As much of the data in educational research is collected using Likert scales, the performance of different estimators under combined conditions of data type and levels of non-normality is an area in need of research. Many multilevel simulation studies only have considered maximum likelihood-based estimation techniques; guidelines for MCFA researchers using ordinal measures is an area in need of attention (Kim et al., 2016). Further, the performance of the partially saturated model level-specific is an area which may be extended. Such investigations may assist MCFA researchers in understanding conditions sensitive to model misspecification and can help inform best practices for using the methodology in applied settings. Finally, study of additional fit indices is warranted. For example, Ene (2020) reported the utility of using Akaike's Information Criteria (AIC) and the Bayesian Information Criteria (BIC) when comparing across complex MCFA structures. Ene (2020) used a correctly specified model in her simulation. Additional study of the AIC and BIC indices under situations where

MCFA models are misspecified at one level, or even across levels, could assist applied researchers.

As an analysis technique, MCFA allows education researchers many advantages. It can accurately accommodate clustered data, provide refined standard errors for parameter testing and interpretation, and allow construction and testing of hypothesized models at different levels. Statistical and computing advances now allow complex procedures to be modeled relatively easily. These advantages can allow for advances in the conceptualization of constructs and theory building at higher analysis levels; however, interpretation and explanation of the MCFA results is needed at all levels so as not to impede future construct validation efforts. We hope that this introduction too provides education researchers additional tools for using and interpreting multilevel confirmatory factor analysis models.

References

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods, 35*(3), 439–460.

Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus.* Technical appendix. Muthén & Muthén.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences.* Guilford Publications.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*(1), 10–17.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research.* Guilford Publications.

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment, 23*(3), 225–241.

DiStefano, C. (2016). Examining fit with structural equation models. In *Principles and methods of test construction. Standards and recent advances* (pp. 166–193). Hogrefe.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *Latent variable growth curve modeling.* Lawrence Erlbaum Associates.

Ene, M. C. (2020). *Investigating accuracy of model fit indices in multilevel confirmatory factor analysis* (Doctoral dissertation, University of South Carolina).

Ene, M., Leighton, E. A., Guo, Z., McGrath, K. V., DiStefano, C., & Monrad, D. M. (2016, April). *Relationships between school climate and indicators of school effectiveness: A multilevel investigation.* Paper presented at the meeting of the American Educational Research Association.

Ene, M., Leighton, E., McGrath, K. V., DiStefano, C., & Monrad, D. M. (2018, April). *School climate and school effectiveness across respondent groups: A multilevel investigation.* Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Ene, M., Leighton, E., Guo, Z., McGrath, K. V., DiStefano, C., & Monrad, D. M. (2017a, April). *School climate and school effectiveness across organizational levels: A multilevel investigation.* Poster presented at the annual meeting of the American Educational Research Association, San Antonio, TX.

Ene, M., Leighton, E., Guo, Z., McGrath, K. V., DiStefano, C., & Monrad, D. M. (2017b, April). *Relationships between school climate and indicators of school effectiveness: A multilevel investigation.* Poster presented at the annual meeting of the American Educational Research Association, Washington, DC.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches.* Guilford Publications.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Erlbaum.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications (Quantitative Methodology Series).* Routledge.

Hox, J. J., & Maas, C. J. (2002). Sample sizes for multilevel modeling. level modeling. In J. Blasius, J. Hox, E. de Leeuw & P. Schmidt (Eds.), *Social science methodology in the new millennium: Proceedings of the fifth In ternational conference on logic and methodology.* http://www.fss.uu.nl/ms/jh/publist/simnorml.pdf

Hsu, H. Y., Kwok, O. M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research, 50*(2), 197–215.

Hu, L. T., & 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.

Huang, F. L., & Cornell, D. G. (2016). Using multilevel factor analysis with clustered data: Investigating the factor structure of the Positive Values Scale. *Journal of Psychoeducational Assessment, 34*(1), 3–14.

Huang, F. L. (2017). *Conducting multilevel confirmatory factor analysis using R* (Unpublished manuscript). http://faculty.missouri.edu/huangf/data/mcfa/MCFA%20in%20R%20HUANG.pdf

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6–23.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8*(3), 325–352.

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research, 51*(6), 881–898.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*, (4th Ed.). Guilford Press.

Konold, T., Cornell, D., Huang, F., Meyer, P., Lacey, A., Nekvasil, E., Heilbrun, A., & Shukla, K. (2014). Multilevel multi-informant structure of the Authoritative School Climate Survey. *School Psychology Quarterly, 29*(3), 238.

Kwon, H. (2011). *A Monte Carlo study of missing data treatments for an incomplete level-2 variable in hierarchical linear models* (Doctoral dissertation). The Ohio State University.

Lai, M. H., & Kwok, O. M. (2015). Examining the rule of thumb of not using multilevel modeling: The "design effect smaller than two" rule. *The Journal of Experimental Education, 83*(3), 423–438.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129–149.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational measurement, 28*(4), 338–354.

Muthén, B. (1993). Latent variable modeling of growth with missing data and multilevel data. In *Multivariate analysis: Future directions, 2* (pp. 199–210).

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376–398.

Muthén, L. K., & Muthén, B. O. (1998–2021). *Mplus user's guide.* (8th Ed.). Muthén & Muthén.

O'Connell, A. A., & McCoach, D. B. (Eds.). (2008). *Multilevel modeling of educational data.* IAP.

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research, 49*(6), 518–543.

Raykov, T., & DiStefano, C. (2021). Design effect in multilevel settings: A commentary on a latent variable modeling procedure for its evaluation. Educational and Psychological Measurement. https://doi.org/10.117700131644211019447

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16*(4), 583–601.

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology, 5*, 81.

Schumacker, E., & Lomax, G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). Routledge.

Stapleton, L. M. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock & R. O. Muller (Eds). *Structural equation modeling: A second course* (2nd ed. , pp. 521–562). IAP.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481–520.

Stockford, S. M. (2009). *Meta-analysis of intraclass correlation coefficients from multilevel models of educational achievement*. Arizona State University.

Zyphur, M. (2019). *Mplus workshop at the university of Melbourne*, February 4–8, 2019.

Christine DiStefano is a Psychometrician and is the E. S. Gambrell Professor of Educational Studies at the University of South Carolina. Her research interests include structural equation modeling with categorical data, classification, and investigations of social-emotional learning in school settings.

Tiejun Zhang is a doctoral student in Educational Studies at the University of South Carolina. His research interests include latent variable analysis and program evaluation in the educational field.

Chapter 3
Multilevel Model Selection: Balancing Model Fit and Adequacy

D. Betsy McCoach, Sarah D. Newton, and Anthony J. Gambino

Abstract Multilevel model (MLM) selection is a consequential and influential decision point in statistical modeling. It dictates how we construct and present knowledge about a phenomenon in our specific field of interest. It also shapes future research, which typically builds upon what we already "know" from existing literature informed by earlier model selection processes. Therefore, competent model evaluation incorporates two main sources of evidence, at minimum: (1) model fit information–the relative degree to which each competing model fits the current data, as evidenced by model fit criteria like the Akaike Information Criterion (AIC; Akaike et al. in Second International Symposium on Information Theory. Academiai Kiado, 1973) and Bayesian Information Criterion (BIC; Schwarz in The Annals of Statistics 6(2):461–464, 1978), and (2) model adequacy data, which indicates the predictive utility of each competing model, or the capacity of the specified set of parameters to explain variance in the outcome of interest. The current chapter briefly reviews MLM estimation techniques; presents several popular model selection criteria; describes an MLM framework for decomposing outcome variance as a measure of model adequacy (Rights and Sterba in Psychological Methods 24:309–338, 2019); details current controversies and concerns for investigating competing MLMs; and provides recommendations for researchers engaged in the MLM selection process.

Keywords Model selection · Akaike Information Criterion · Bayesian Information Criterion · Model fit · Model Adequacy · Variance Decomposition

3.1 Introduction

How do researchers evaluate and choose one of several competing multilevel models (MLMs)? How do they determine the utility of a given model, especially compared to other plausible data representations? It is important to consider both model fit (use of model selection criteria to choose among competing models) and

D. B. McCoach (✉) · S. D. Newton · A. J. Gambino
Department of Educational Psychology, University of Connecticut, Storrs, CT, USA
e-mail: betsy.mccoach@uconn.edu

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_3

model adequacy (the ability of predictors to explain outcome variability). As such, model selection processes must assess both fit and adequacy for each model under consideration.

In this chapter, we describe this model evaluation process. After providing a brief conceptual overview of MLM estimation, we identify common measures of model fit and adequacy; highlight several areas of controversy/confusion; and provide general recommendations for evaluating MLM fit and adequacy. Related to fit, we review the concept of deviance and explain how to use the chi-squared difference test to compare the deviances of two nested models. We also describe index comparison approaches (e.g., the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to model evaluation). Then we consider model adequacy, reviewing current proportion of variance explained-type measures to determine the predictive power of MLMs. Finally, we offer guidance for assessing model fit and adequacy in MLM.

3.2 Conceptual Overview of Estimation in MLM

Before proceeding, it is useful to have a rudimentary conceptual understanding of estimation in MLM.[1] To keep things simple, we contextualize this discussion using the unconditional random effects model with a single, continuous outcome variable, $Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$. In this model: Y_{ij} represents the predicted value on the outcome for person i in cluster j; γ_{00} denotes the overall intercept—the expected person-level value on the outcome, absent any additional person- or cluster-level information; u_{0j} conveys unexplained variance in the outcome across clusters; and e_{ij} indicates the degree of person-level divergence from the overall intercept.

In non-clustered data, the sample mean provides our “best guess” about the population mean (the “expected” value in the population). But with clustered data, what is the expected value of the outcome variable, γ_{00}? Imagine randomly sampling students from 100 schools. Further, the sample sizes within the schools vary widely: the smallest cluster size is 2 and the largest cluster size is 1000. How should we determine expected achievement? One option would be to ignore clustering and to take the sample mean, such that every person is weighted equally; however, this means that schools with more students would have more influence on the expected mean than smaller schools. Alternatively, we could compute the mean of school means. Yet weighting all schools equally gives small schools a disproportionately large influence on the expected mean.[2]

In MLM, larger clusters generally exert more influence on the expected mean. However, the intraclass correlation coefficient (ICC) tempers that effect. In the

[1] In this chapter, we focus exclusively on maximum likelihood estimation, but it is also possible to use fully Bayesian methods to estimate MLMs.

[2] In addition, the school mean that is computed from a school with 1,000 students is likely to be a much better estimate of the school’s performance than a school mean that is computed from only two students, a point to which we return when we discuss Empirical Bayes estimates.

unconditional random effects ANOVA model, the ICC represents the proportion of between-cluster variance in the outcome variable, indicating the degree of dependence within a cluster:

$$\mathrm{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{3.1}$$

Where τ_{00} conveys between-cluster variability, σ^2 denotes within-cluster variability, and their sum (in the denominator) represents total variance in the outcome.

An ICC of 0 suggests that two randomly selected people within a given cluster are no more similar to each other than two randomly selected people from different clusters. So, ignoring clustering (and treating every unit in the sample as independent) seems reasonable. When the ICC is 1, each cluster member is basically a complete replicate of every other member of the same cluster on the outcome. Hence, sampling within a cluster would be unnecessary because there is no within-cluster variability—every member of the cluster shares the same value on the outcome of interest. So, calculating the mean of cluster means might be more useful (given the deterministic nature of within-cluster performance). When the ICC is 0, the influence of each cluster on γ_{00} is determined by the sample size of the cluster. When the ICC is 1, each cluster has an equal influence on γ_{00}, regardless of its size (Snijders & Bosker, 2012). In reality, the ICC is typically between 0 and 1. Therefore, γ_{00} is a compromise between the sample mean (as it would be when ICC = 0) and the mean of cluster means (as it would be when ICC = 1). The higher the ICC, the more γ_{00} approaches the mean of cluster means; the lower the ICC, the more γ_{00} approaches the sample mean.

3.3 Reliability of Cluster *J*

Potential expected values of the *true* cluster mean include estimates of γ_{00} and $\overline{Y}_{.j}$. Empirical Bayes estimation combines these values, based on the reliability of cluster *j*, which incorporates three pieces of information: within-cluster variability (σ^2), between-cluster variability (τ_{00}), and the number of observations per cluster, n_j [see McCoach et al. (2022) for more details about MLM estimation].

$$\text{Reliability of } \hat{\beta}_{0j} = \frac{\tau_{00}}{\tau_{00}+\frac{\sigma^2}{n_j}} \tag{3.2}$$

Each cluster has its own estimate of reliability, whereas variance estimates (τ_{00}, σ^2) remain constant across clusters. Theoretically, the reliability of cluster *j* can range from 0 (when there is no between-cluster variability in the outcome) to 1 (when there is no within-cluster variability in the outcome). However, within a given sample, the lower bound for reliability for any given cluster is the ICC, which occurs when the cluster has only one unit ($n_j = 1$).

The ICC features prominently in this reliability formula (Raudenbush & Bryk, 2002). Larger ICCs, which indicate that within-cluster variance is small relative to between-cluster variance, produce higher reliability. In other words, reliability is higher when cluster means vary substantially across level-2 units (holding cluster size constant). In addition, for a given pair of between- and within-school variance values, larger cluster sizes (n_j's) result in higher reliability. So, increasing group size, increasing homogeneity within clusters, and increasing heterogeneity between clusters all increase reliability. When the reliability in cluster j is higher, more weight is placed on the sample mean as the estimate of the true school mean. Conversely, when the reliability of the cluster j is lower, more weight is placed on the estimate of γ_{00} as the expected true school mean.

3.4 Maximum Likelihood Estimation

We often use maximum likelihood (ML) estimation for MLM. The goal of ML estimation is to find the set of parameter values that maximizes the likelihood of observing the actual data. The most-common techniques for estimating variance components for MLMs with normal response variables are full-information maximum likelihood (FIML) and restricted maximum likelihood (REML) estimation.

3.4.1 FIML

In FIML, the estimates of the variance and covariance components are conditional upon the point estimates of the fixed effects. This method chooses estimates of **Γ** (i.e., the fixed-effect variance/covariance matrix), **T** (i.e., the random effect variance/covariance matrix; see Eq. 3), and σ^2 “that maximize the joint likelihood of these parameters for a fixed value of the sample data, Y” (Raudenbush & Bryk, 2002, p. 52). Thus, the number of model parameters includes both fixed effects and variance/covariance components.

$$T = \begin{bmatrix} \tau_{00} & & & & \\ \tau_{01} & \tau_{11} & & & \\ \tau_{02} & \tau_{12} & \tau_{22} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \tau_{0q} & \tau_{1q} & \tau_{2q} & \ldots & \tau_{qq} \end{bmatrix} \tag{3.3}$$

3.4.2 REML

In contrast, REML maximizes the joint likelihood of $\mathbf{T}$ and σ^2 given the observed sample data, $\mathbf{Y}$. Therefore, when estimating variance components, REML takes the uncertainty due to loss of degrees of freedom from estimating fixed parameters into account, whereas FIML does not (Goldstein, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Hence, under REML, the number of reported estimated parameters includes only the variance and covariance components.

3.4.3 FIML vs. REML

When comparing nested models under REML, the fixed-effects structure must be consistent across the null and alternative models. To compare models that differ in both fixed and random effects requires FIML (Goldstein, 2010; McCoach & Black, 2008; McCoach et al., 2018; Snijders & Bosker, 2012).

When the number of clusters (level-2 units) is large, REML and FIML produce similar estimates of the variance components. However, for small numbers of clusters, FIML tends to underestimate variance components, and REML results may be more realistic (Raudenbush & Bryk, 2002). A simple formula, (J–F)/J, where J is the number of clusters and F is the number of fixed effects in the model, provides a rough approximation of the degree of downward bias in FIML estimates (Raudenbush & Bryk, 2002). For example, when fitting a model with three fixed effects in a sample containing observations from 20 clusters, the level-2 variance components are likely to be 75% (((20−5)/20) = 0.75) as large in FIML as they are in REML. In contrast, fitting a model with 5 fixed effects and 500 clusters results in very little bias (500–5/500 = 0.99).

3.5 Model Selection

Models simply represent approximations of reality (Burnham & Anderson, 2004). If we accept this basic premise, how should we select from a set of competing models? Guided by theory and informed by data, balancing parsimony and model complexity, the ideal model should adequately describe the observed data to a satisfactory extent, but exclude unnecessary complications (Snijders & Bosker, 2012). Hence, “the best model is the one that provides an adequate account of the data while using a minimum number of parameters” (Wagenmakers & Farrell, 2004).

In pursuit of this overarching goal, three general principles should guide the model selection process (Burnham & Anderson, 2004): (1) Parsimony—estimating more parameters cannot lead to worse model fit (Forster, 2000), but does the improvement justify inclusion of additional parameters? (2) Comparison of plausible competing hypotheses—why compare a specified model to the often-implausible null model (Burnham & Anderson, 2004) in which most substantive parameters are constrained to be exactly zero in the population? (3) Strength of support—we should gauge *support for* a specified theoretical model (Burnham & Anderson, 2004) rather than *evidence against* the atheoretical null model.

3.6 Criteria for Evaluating Model Fit

There are a variety of options for examining and comparing the model/data fit for a set of competing MLMs. We can compare *nested* models using the Likelihood Ratio Test (i.e., LRT, Deviance Difference Test) and either *nested* or *non-nested* models with information criteria (ICs), like the AIC and BIC. Two models are considered hierarchically nested when "the more complex model includes all of the parameters of the simpler model plus one or more additional parameters" (Raudenbush et al., 2000, pp. 80–81).

3.6.1 Likelihood Ratio Test (LRT)

We calculate the deviance as: -2 times the difference between the log-likelihood of the specified model and the log-likelihood of a saturated model that fits the sample data perfectly (i.e., -2LL); higher deviances indicate greater model misfit (Singer & Willett, 2003). The LRT compares the deviances from competing nested models. But we can only interpret these deviance differences if both models: (1) are hierarchically nested; (2) analyze the same observations; and (3) use FIML estimation (if the models differ in their fixed-effects structure). The LRT's null hypothesis is that the restricted/constrained null model (M_0, which estimates fewer parameters) is correct; the alternative hypothesis posits that the unrestricted/unconstrained, more-complex alternative model (M_1) is true (Dominicus et al., 2006).

In sufficiently large samples, under standard normal theory assumptions, using the same set of observations, the difference between the deviances of two hierarchically nested models follows an approximate χ^2 distribution with degrees of freedom equal to the difference in the number of parameters estimated between the two candidate models (Raudenbush & Bryk, 2002; Singer & Willett, 2003).

The deviance of the simpler model (D_0) is based on p_0 estimated parameters, whereas the deviance of the more-complex model (D_1) is associated with p_1 estimated parameters. Because the simpler model has fewer parameters ($\mathrm{p}_0 < \mathrm{p}_1$), the deviance of the simpler model must be at least as large as the deviance of the more-parameterized model ($D_0 \geq D_1$).

The LRT compares the difference in model deviances ($\Delta D = D_0 - D_1$) to the critical value of χ^2, with degrees of freedom equal to the difference in the number of estimated parameters ($\Delta \mathrm{p} = \mathrm{p}_1 - \mathrm{p}_0$). If the more-complex model fails to reduce the deviance by a substantial amount, we retain the simpler model (M_0). However, when the change in deviance (ΔD) exceeds the critical value of χ^2 with p_1–p_0 degrees of freedom, the additional parameters result in statistically significantly improved model fit, which favors the more-parameterized model (M_1). Therefore, a statistically significant decrease in the deviance favors the more-complex model (M_1), whereas a non-significant decrease favors the less-complex model (M_0; McCoach & Black, 2008; Raudenbush & Bryk, 2002).

LRTs for Random Effects. Evaluating variance components is trickier. Traditional LRTs are less appropriate for testing random effects (Berkhof & Snijders, 2001; Raudenbush & Bryk, 2002) because the variance components are boundary parameters—they cannot be normally distributed around a mean of zero if the null hypothesis is true (Dominicus et al., 2006; Stoel et al., 2006). In other words, because variance components cannot be negative, their sampling distributions may only contain positive values and zero; these distributions must be bounded at zero. Therefore, using two-tailed tests to evaluate variances is inappropriate (Raudenbush & Bryk, 2002). Computing standard errors around variance estimates also becomes more problematic as variance components approach zero (Baayen et al., 2008), and using confidence intervals and other statistical tests based on typical standard errors is questionable. The unadjusted LRT is too conservative to test random effects (Self & Liang, 1987; Stoel et al., 2006; Stram & Lee, 1994); *p*-values are larger than they should be (Baayen et al., 2008; Dominicus et al., 2006; Stoel et al., 2006) resulting in elevated Type II error rates (i.e., researchers failing to reject the null hypothesis when it is false; Dominicus et al., 2006; Stoel et al., 2006) and decreased power (Berkhof & Snijders, 2001; Stoel et al., 2006), which may increase the likelihood of constraining slope variances that should randomly vary.

Hence, we typically use a modified χ^2 sampling distribution to test boundary parameters. This resembles a mixture of zero and the appropriate positive χ^2 distribution. In other words, when investigating a single boundary parameter, 50% of the variance component's sampling distribution indicates the possibility of zero variance, and 50% of the distribution represents positive variance values (Berkhof & Snijders, 2001; Miller, 1977; Self & Liang, 1987; Snijders & Bosker, 2012; Stoel et al., 2006). See McCoach et al. (2022) for a more-detailed description of mixed χ^2 distributions.

3.6.2 Recommendations for Testing Random Effects

When comparing two models that differ by only one fixed effect (no random effects), the χ^2 critical value is 3.841 at $\alpha = .05$. However, if models differ by a single random effect, we instead conduct a one-tailed hypothesis test using the χ^2 critical value for $p = 0.10$ to test for statistical significance at alpha $= 0.05$ (Snijders & Bosker, 2012). For this desired alpha level, the critical value of χ^2 with one degree of freedom is 2.706. So, if two models differ by one parameter, and it is a variance component, we compare our test statistic to $\chi^2 = 2.706$. Snijders and Bosker (2012) present a more-detailed discussion of this issue and display a table with the correct critical values for the comparison of models that differ by one or more randomly varying slopes. For models differing in both fixed and random effects, the proper LRT distribution depends on the number of boundary parameters of interest (k, with $k + 1$ distributions combined to formulate the appropriate mixture distribution) and the number of unconstrained parameters of interest (u, with u degrees of freedom for the first distribution in the mixture; Stoel et al., 2006).

Another complication of testing random effects is that some statistical packages (like HLM, M*plus*, and Stata) produce and report standard errors for variance components, some (like R) do not, and some offer user-specified options to obtain such estimates. For example, SAS includes the *covtest* option in Proc Mixed, which produces standard errors and statistical tests for covariance estimates (SAS Institute Inc., 2020). Hence, it is important to carefully select an MLM package and understand its default options.

In addition, the HLM software provides univariate χ^2 tests to investigate the null hypotheses that level-2 (intercept and slope) variances are 0. Statistically significant χ^2 values suggest retaining these variance components; non-statistically significant χ^2s suggest eliminating them. However, Raudenbush and Bryk (2002) caution that these χ^2 tests provide approximate, not exact, probability values because they are simple univariate tests estimated only for clusters with sufficient data to compute separate OLS regressions.

3.6.3 Other Issues with Modeling Random Slopes

In addition to boundary issues, there are several other things to consider when modeling random slopes. First, near-zero random effects (i.e., slope variances not significantly different from 0) may cause the model to iterate repeatedly, decreasing computational speed. Or the model may fail to converge altogether (McCoach et al., 2018; Palardy, 2011). This “failure to converge is not due to the defects of the estimation algorithm” (Bates et al., 2015, p. 19), but is “a straightforward consequence of attempting to fit a model that is too complex to be supported by the data” (Bates et al., 2015, p. 19). One solution is to fit an MLM with a more-parsimonious variance/covariance structure (Bates et al., 2015). Interestingly though, an MLM that fails to converge in one statistical package may converge in another package (McCoach et al., 2018).

Furthermore, in unstructured **T** matrices, estimating additional random slopes increases the complexity of the model more than adding fixed effects. Note that the variance component structure in an estimated MLM represents a partitioning of the residual variance (i.e., unexplained variance in the outcome). So, estimating additional variance components increases the computational load of model estimation as the unstructured **T** matrix includes the random slope *variance*s as well as the *covariances* between each new random slope and each existing random intercept/slope in the model. Therefore, the number of estimated parameters increases by the number of random slopes we add to our model, as well as the number of covariances between those new random slopes and all existing variance components. This, in turn, increases computational burden and model estimation time. Therefore, it is important to model variance components judiciously. Incorporate theoretically defensible random effects into MLMs but resist the temptation to include all possible random effects by default, without carefully considering their purpose and necessity. Just because it is possible to specify a random slope does not mean that it is advisable (McCoach & Cintron, 2021).

3.6.4 Information Criteria (ICs)

Fit criteria provide another tool for evaluating and comparing sets of competing models (Konishi & Kitagawa, 2008) by offering a formal method for quantifying the strength of data-based evidence for, and ranking of, each competing hypothesis (Burnham et al., 2011). There are several advantages to consulting ICs during model selection, rather than relying solely upon deviance statistics and χ^2 difference tests to evaluate the goodness of fit of MLMs. First, ICs such as the AIC (Akaike, 1973) and BIC (Schwarz, 1978) allow for comparison of non-nested models, assuming FIML estimation and a consistent sample across models. Furthermore, ICs quantify the degree to which the given model represents an improvement over comparison models.

Information criteria (IC) generally feature two components: the deviance (which conveys the fit of the estimated model) and some penalty on the deviance for each additional estimated parameter (Dominicus et al., 2006; Lee & Ghosh, 2009; Zucchini et al., 2011). Although the number of estimated model parameters changes across models, the per-parameter penalties of the AIC and BIC remain constant for a given sample size. When using ICs to compare two models, we favor the model with the lower IC. Table 3.1 enumerates formulae for several popular ICs used today.

Sample Size for Computing the BIC in MLM. Whereas the AIC does not explicitly incorporate sample size into its computational formula, other ICs like the BIC do. In single-level models, we assume complete independence of observations under simple random sampling, so sample size determination is straightforward. However, for MLMs, we have (by definition) multiple sample sizes to choose. So, selecting the correct sample size for BIC computation is more complicated—should we use the total level-1 sample size or the number of clusters at our highest level?

In clustered designs, we could also employ the effective sample size (e.g., $N_{effective}$), which adjusts the total number of level-1 units for the two-stage sampling procedure used to collect multilevel data (Snijders & Bosker, 2012). It accounts for both the homogeneity within clusters (ICC) and the average cluster size ($\overline{n}_{.j}$):

$$n_{eff} = \frac{N}{DEFF} = \frac{N}{1+\rho\left(\overline{n}_j - 1\right)} \tag{3.4}$$

Table 3.1 Common information criteria

Information criterion	Definition
Akaike Information Criterion (Akaike, 1973)	$\text{AIC} = -2LL + 2p$
Bayesian Information Criterion (Schwarz, 1978)	$\text{BIC} = -2LL + \ln(n) * p$
Sample size-adjusted Bayesian Information Criterion (SABIC; Sclove, 1987)	$\text{SABIC} = -2LL + \ln((n + 2)/24) * p$
Hannan and Quinn Information Criterion (HQIC, Hannan & Quinn, 1979)	$\text{HQIC} = -2LL + 2 * p * \ln(\ln(n))$
Consistent AIC (CAIC; Bozdogan, 1987)	$\text{CAIC} = -2LL + [\ln(n) + 1] * p$
Finite sample corrected AIC (AICC, Hurvich & Tsai, 1989)	$\text{AICC} = -2LL + (2 * p * n)/(n - p - 1)$

Note In the equations above, n represents the sample size; p denotes the number of estimated model parameters

where $\overline{n}_j$ is the average cluster size, ρ is the ICC, and N is the total sample size (Snijders & Bosker, 2012).

Given multiple conceptualizations of sample size within a multilevel framework, there is no definitive consensus as to which sample size is most appropriate for computing the BIC (Skrondal & Rabe-Hesketh, 2004). Furthermore, different software packages calculate the BIC differently. See McCoach et al. (2022) for more on how different statistical packages compute the BIC differently.

Comparing the AIC and BIC. Notably, the AIC tends to favor more-complex models (Bozdogan, 1987; Whittaker & Stapleton, 2006; Whittaker et al., 2012, 2013) than other available model fit indices (such as the BIC or LRT) because of its small, 2-point, per-parameter penalty. For example, the critical value of χ^2 with one degree of freedom at $\alpha = 0.05$ is 3.841 (or 2.706 for a 1-*df* change involving a variance). When comparing two models that differ by a single degree of freedom, the LRT actually imposes a more-stringent penalty for rejecting the simpler model. In fact, this is true for comparisons of models that differ by up to seven parameters[3]: using the LRT results in an equivalent or more-parsimonious model than the AIC. Conversely, when comparing models that differ by more than seven parameters, the AIC favors more-parsimonious models than the LRT.

Comparatively, the per-parameter penalty for the BIC is often larger than that of the AIC; the sample size must be less than eight for the BIC's per-parameter penalty to drop below AIC's 2-point penalty (Claeskens & Hjort, 2008; Schwarz, 1978). So,

[3] The number seven assumes that we are using the standard critical values for chi-squared with alpha = .05, not critical values that have been adjusted for boundary issues in the variances. With a difference of seven parameters, the traditional, unadjusted LRT imposes a penalty of 14.07 points on the deviance, whereas the AIC features a 14.00-point penalty. However, with a difference of eight parameters, the traditional LRT's critical value is 15.51, whereas the AIC's penalty is 16.00 points. Therefore, for a difference of up to seven parameters, the traditional LRT imposes a larger penalty on the deviance than does the AIC; in contrast, for eight or more additional estimated parameters, the AIC's penalty exceeds that of the unadjusted LRT.

when the AIC favors the simpler model, the BIC also favors this model. And when the BIC favors the more-complex model, the AIC does, too.

But the AIC and BIC do not just differ in per-parameter penalties. The purposes of the AIC and BIC are also fundamentally different, a point that is often ignored in practice. Whereas the AIC "is designed to find the best approximating model to the unknown true data-generating model, the BIC is designed to find the most probable model given the data" (Vallejo et al., 2011, p. 22). Because the AIC and BIC take different approaches to model selection, using these ICs may result in different model selection conclusions.

Additionally, modeling objectives influence the choice of IC. For example, predictive models often include as many explanatory variables as possible, no matter how little each variable contributes. In such a situation, using the AIC[4] might be preferable because of its low threshold for estimating additional parameters. In contrast, when building explanatory models, researchers covet simplicity, including only the most-influential predictors to explain the outcome. Given that the BIC focuses on parsimony by allowing fewer predictors to remain in the model, this IC may be preferable for such a purpose (McCoach et al., 2022).

What Happens When Model Fit Criteria Disagree? The AIC, BIC, and LRT may favor different models. For example, imagine if two models differed by one fixed effect, with a total sample size of 1000 people, nested within 50 clusters. The deviance must decrease by 2.00 points for the AIC, 3.91 points for the BIC-L2 (i.e., BIC calculated with the number of level-2 units), and 6.91 points for the BIC-L1 (i.e., BIC calculated with the number of level-1 units) to favor the more-parameterized model. Given the gap in their penalties, these ICs do not always agree on which model to favor (McCoach & Cintron, 2021).

Hence, in addition to examining model fit criteria, we recommend investigating measures of model adequacy, such as those presented in Rights and Sterba's (2019) MLM variance decomposition framework (see next section). Model adequacy measures can bolster a theory-driven argument for retaining/eliminating specific parameters/models from consideration and provide insight into the necessity of a given parameter by addressing the following:

1. What proportion of within, between, and/or total variance does a specific parameter explain?
2. Does removing this parameter substantially reduce the predictive ability of the model?

Ultimately, researchers must decide which error would be more serious: omitting a potentially necessary parameter or including an unnecessary parameter. In general, we tend to favor retaining potentially important fixed effects and eliminating unnecessary random effects. However, the field of study and specific research context, problem, and questions should inform this choice. Next, we introduce methods for quantifying explained variance in MLM.

[4] Akaike (1973) developed the AIC with a focus on prediction—he saw the ability to forecast future outcomes as the function and utility of statistical models.

3.7 Model Adequacy

In MLM, we generally evaluate model fit relative to other competing models. We can also evaluate model adequacy (i.e., explanatory power of the model—does the model do a good job of explaining scores on the outcome variable?). We can assess model adequacy both relative to competing models and in an absolute sense.

In single-level regression models, an important determinant of the utility of a model is the proportion of outcome variance explained by the model, or R_2. In MLM, the outcome variance is decomposed into multiple pieces called variance components, which exist at each level of the MLM. In addition, in random-coefficients models, the relationships between the dependent variable and level-1 independent variables vary as a function of the level-2 unit or cluster variable. Therefore, quantification of the variance explained by a given set of predictors is much more complicated in MLM than in single-level regression.

Additionally, deciding how to compute and report variance-explained measures in MLM requires explicit consideration of the context and goals of the research. Conceptually, we may be interested in measuring variance explained within clusters, variance explained between clusters, and/or total variance explained (both within and between clusters). Cluster-level (level-2) variables cannot explain within-cluster variance, but they can explain between-cluster variance. And group-mean centered level-1 (within-cluster) variables can explain within-cluster, but not between-cluster, variance. However, both types of variables also account for some portion of the total outcome variance through the level-specific variance they explain. Imagine a situation in which 5% of the outcome variance lies between clusters and 95% of the outcome variance lies within clusters. A variable that explains 80% of the between-cluster variance only explains 4% (80%*5%) of the total variance. In contrast, a variable that explains 5% of the within-cluster variance explains 4.75% of the total variance for the specified model.

3.7.1 Proportional Reduction in Variance Statistics

The most-commonly reported multilevel pseudo-R^2 statistics (i.e., measures of the proportion of outcome variance explained by a given model) are Raudenbush and Bryk's (2002) *proportional reduction in variance* statistics, which compare the residual variance from the full (more-parameterized) model to the residual variance from a simpler "base" model. If the full model explains additional variance, then the residual variance should decrease. In such a scenario, the full residual variance for the baseline model should be greater than the residual variance for the full model. See Raudenbush and Bryk (2002) and McCoach et al. (2022) for additional details about how to compute *proportional reduction in variance* statistics at different MLM levels.

3.7.2 Variance Decomposition Framework for MLM

More recently, Rights and Sterba (2019) developed "an integrative framework of R-squared measures for MLMs with random intercepts and/or slopes based on a completely full decomposition of variance" (p. 309). This framework allows for computation of existing variance-explained measures and alleviates the need to estimate multiple models.

However, partitioning total outcome variance into within- and between-cluster variance requires the group-mean centering of all level-1 predictors. To retain the cluster-level variance from each level-1 predictor, we re-introduce the aggregates of the level-1 variables into the level-2 model.[5] If we decompose all predictors into within- and between-cluster variables, then the model-implied total outcome variance is attributable to five specific sources of variation: level-1 predictors via fixed slopes (f_1); level-2 predictors via fixed slopes (f_2); level-1 predictors via random slope variation/covariation (v); cluster-specific outcome means via random intercept variation (m)[6]; and level-1 residuals (σ^2). Decomposing the model-implied total variance in this way enables the computation of multiple variance-explained measures,[7] each of which provides insights into the model's predictive capability (Rights & Sterba, 2019).

Careful examination of this framework reveals some important observations. First, assuming group-mean centered level-1 variables, three sources of variation contain only within-cluster variance: level-1 predictors via fixed slopes (f_1); level-1 predictors via random slope variation/covariation (v); and level-1 residuals (σ^2). Therefore, we can evaluate the proportion of within-cluster variance explained by the specified set of level-1 predictors (f_1) as well as the variances and covariances of their associated randomly varying slopes (v). We can also distinguish these sources of variance from the level-1 residual variance. Second, the remaining sources of variation (i.e., level-2 predictors via fixed slopes (f_2) and cluster-specific outcome means via random intercept variation (m)) contain only between-cluster variance. In other words, between-cluster variance is either explained by our set of level-2 predictors (f_2) or random variation in the intercept (m).

Understanding Proportion of Variance-explained Measures. Calculation of the R^2 measures described in Rights and Sterba (2019) requires model parameter

[5] There is one exception: if the level-1 variable has only within-cluster variance, and has no between-cluster variance, then this is not necessary. For example, if each of the clusters contained an even number of people who were randomly assigned to either the treatment or control group, then regardless of coding/centering ([0, 1] dummy coding or [−1, 1] effect coding), the cluster mean for every cluster is identical. In such a scenario, it is unnecessary to add the aggregate back in at level 2. In fact, in such a situation, doing so would be problematic, given the complete lack of variability in the cluster means (The cluster means would be constant, not variable across all clusters). Using an effect code ([+1, −1] or [+1/2, −1/2]) produces the same result as group-mean centering in this scenario.

[6] m is technically $m = \tau_{00}$ as long as the level-1 predictors all have means of 0.

[7] The authors showed correspondence between their integrative framework and other R^2 measures commonly used in MLM. See Rights and Sterba (2019) for more details.

estimates (i.e., level-1 residual variance, fixed-effect estimates for all predictors, variance components for all random intercept/slope parameters, and covariances among these variance components), as well as the sample variance–covariance matrix for the set of predictors included in the MLM. Given this information, and under the assumption that all level-1 variables are group-mean centered (including categorical variables), the following formulae produce Rights and Sterba's proposed R^2-measure components. The f_1 component (i.e., variance explained by fixed level-1 predictors) is

$$f_1 = \boldsymbol{\gamma}'_{\mathrm{W}} \Phi \boldsymbol{\gamma}_{W}, \tag{3.5}$$

where $\boldsymbol{\gamma}_{\mathrm{w}}$ is the vector of level-1 fixed effect estimates (excluding the intercept term), and Φ_{w} is the variance–covariance matrix for the level-1 predictors. This must include any within-cluster or cross-level interaction variables (Rights & Sterba, 2019).

The f_2 component (i.e., variance explained by the fixed level-2 predictors) is

$$f_2 = \boldsymbol{\gamma}'_{\mathrm{B}} \Phi_{\mathrm{B}} \boldsymbol{\gamma}_{\mathrm{B},} \tag{3.6}$$

where $\boldsymbol{\gamma}_{\mathrm{B}}$ is the vector of level-2 fixed effect estimates, and Φ_{B} is the variance–covariance matrix for the level-2 predictors. This set of level-2 predictors must include any cluster-level interaction product variables.

The v component (i.e., variance explained by random slope variation/covariation) is

$$v = tr(\mathrm{T}\Sigma), \tag{3.7}$$

where *tr*() is the trace function (the sum of the diagonal elements of a matrix),**T** is the random effect variance–covariance matrix (the "**T** au matrix"), and Σ is the variance–covariance matrix of the level-1 predictors with randomly varying slopes. (This includes a variance of 0 for the intercept and covariances of 0 between the intercept and each of the predictors because the intercept is a constant).

The final two components do not require any additional computation–the *m* component is the estimate of random intercept variance (i.e.,τ_{00}), and σ^2 is the level-1 residual variance estimate.

The model-implied level-1 outcome variance is the sum of the f_1, *v*, and σ^2 components; the model-implied level-2 outcome variance is the sum of the f_2 and *m* components; and the model-implied total outcome variance is the sum of all five components (Rights & Sterba, 2019).

Notably, group-mean centering every level-1 variable allows the contribution of fixed effects to be partitioned into two components, f_1 and f_2, which allows for computation of the within, between, and total R^2 measures. When modeling non-group-mean centered level-1 variables, it is still possible to compute total R^2 measures. However, it is not possible to decompose f into f_1 and f_2; therefore, it is not possible to compute separate within- and between-level R^2 measures. In this case, the overall f component represents variance attributable to all predictors via fixed slopes:

$$f = \boldsymbol{\gamma}'\Phi\gamma, \tag{3.8}$$

where $\boldsymbol{\gamma}$ is the vector of fixed-effect estimates for all predictors (excluding the intercept) and Φ is the variance–covariance matrix for all predictors. Additionally, if there are any non-group-mean-centered level-1 variables, the m component is:

$$m = \boldsymbol{\mu}'T\mu, \tag{3.9}$$

where $\boldsymbol{\mu}$ is a column vector containing the means of the predictors with random slopes (including "1" as its first element to represent the random intercept) (Rights & Sterba, 2019).

Using these five main sources of variation (i.e., f_1, f_2, v, m, σ^2), we can compute several variance-explained measures. For example, the proportion of within-cluster outcome variance explained by level-1 predictors via fixed slopes ($R_w^{2(f_1)}$) indicates how well the *fixed effects* for the set of level-1 predictors (slopes) explain within-cluster variance in the outcome variable:

$$R_w^{2(f_1)} = \frac{f_1}{f_1+v+\sigma^2} \tag{3.10}$$

If any level-1 slopes randomly vary across clusters, we compute the proportion of within-cluster variance explained by level-1 predictors via random-slope variation/covariation:

$$R_w^{2(v)} = \frac{v}{f_1+v+\sigma^2} \tag{3.11}$$

This indicates the proportion of within-cluster variance explained by allowing slopes to randomly vary across clusters. Adding cross-level interactions to explain between-cluster variability in the slopes decreases v and increases f_2. Therefore, $R_w^{2(v)}$ is the slope variance that is not explained by level-2 (between-cluster) variables.

Inspecting $R_w^{2(v)}$ also provides insight into the pervasive question: should all, some, or none of the slopes in my MLM randomly vary? To address this issue, we can examine the $R_w^{2(v)}$ separately for each random slope to determine how much within-cluster variance each slope explains. Specifically, we fit a model in which only one slope randomly varies. The $R_w^{2(v)}$ for the model with a single randomly varying slope indicates the proportion of within-cluster variance attributable to that random effect.

The total proportion of within-cluster outcome variance explained by level-1 predictors via fixed slopes and random slope variation/covariation ($R_w^{2(f_1v)}$) is the sum of the variance attributable to level-1 predictors via fixed slopes (f_1) and random-slope variation/covariation (v), divided by the total level-1 variance:

$$R_w^{2(f_1v)} = \frac{f_1+v}{f_1+v+\sigma^2} = 1 - \frac{\sigma^2}{f_1+v+\sigma^2} \tag{3.12}$$

This is analogous to computing Raudenbush and Bryk's (2002) proportion reduction in level-1 residual variance. One minus $R_w^{2(f_1 v)}$ is the proportion of unexplained within-cluster residual variance (i.e., variance not explained by either the fixed effects or the randomly varying slopes for the specified set of level-1 predictors).

$$1 - R_w^{2(f_1 v)} = \frac{\sigma^2}{f_1+v+\sigma^2} \tag{3.13}$$

In addition, at level 2, the proportion of between-cluster outcome variance explained by level-2 predictors via fixed slopes ($R_b^{2(f_2)}$) captures the degree to which the set of level-2 predictors explains between-cluster variance in the outcome variable.

$$R_b^{2(f_2)} = \frac{f_2}{f_2+m} = \frac{f_2}{f_2+\tau_{00(f)}}, \tag{3.14}$$

This is equivalent to Raudenbush and Bryk's proportion reduction in level-2 variance.

Note that one minus the proportion of between-cluster outcome variance explained by level-2 predictors via fixed slopes is the proportion of between-cluster variance that remains unexplained by the model—the proportion of level-2 residual variance in the intercept:

$$R_b^{2(m)} = 1 - \frac{f_2}{f_2+m} = \frac{m}{f_2+m} = \frac{\tau_{00(f)}}{f_2+\tau_{00(f)}} \tag{3.15}$$

Furthermore, the proportion of total variance explained by cluster-specific outcome means via random intercept variation ($R_t^{2(m)}$) is the conditional ICC. The conditional ICC indicates the proportion of residual between-cluster variance in the intercept, after accounting for all of the variables in the model.[8] It indicates how much variance is explained by allowing the intercept to randomly vary across clusters (Rights & Sterba, 2019). Specifically, $R_t^{2(m)}$ is

$$R_t^{2(m)} = \frac{m}{m+f_2+f_1+v+\sigma^2} = \frac{\tau_{00(f)}}{\tau_{00(f)}+f_2+f_1+v+\sigma^2} \tag{3.16}$$

Notably, the proportion of total variance explained by level-1 and level-2 fixed effects ($R_t^{2(f)}$) is somewhat analogous to an R-squared measure in single-level regression. It captures the proportion of outcome variance explained by fixed effects across levels of the MLM (Rights & Sterba, 2019):

$$R_t^{2(f)} = \frac{f_1+f_2}{f_1+f_2+v+m+\sigma^2} \tag{3.17}$$

We can also decompose $R_t^{2(f)}$ into its two constituent pieces, $R_t^{2(f_1)}$ and $R_t^{2(f_2)}$, where $R_t^{2(f_1)}$ is the proportion of total variance explained by level-1 slopes and cross-level interactions (given that all level-1 predictors are group-mean centered) (Rights & Sterba, 2019):

[8] Note: If both v and m were 0, there would be no need for an MLM—we would require only one residual component, σ^2.

$$R_t^{2(f_1)} = \frac{f_1}{f_1+f_2+v+m+\sigma^2} \tag{3.18}$$

And ($R_t^{2(f_2)}$) is the proportion of total variance explained by level-2 predictors:

$$R_t^{2(f_2)} = \frac{f_2}{f_1+f_2+v+m+\sigma^2} \tag{3.19}$$

Finally, we may wish to report the proportion of outcome variance explained by the variances and covariances of our model's randomly varying slopes ($R_t^{2(v)}$):

$$R_t^{2(v)} = \frac{v}{f_1+f_2+v+m+\sigma^2} \tag{3.20}$$

See McCoach et al. (2022) for a more in-depth discussion of these R^2 measures that includes a tutorial on how to compute these measures using R statistical software.

3.8 Conclusion

Measures of model fit and model adequacy provide useful tools for comparing, selecting, and interpreting MLMs in applied research contexts. When used together, they offer distinct, but complementary, perspectives on the utility of theoretical models, allowing researchers to make informed and justifiable modeling decisions. In addition, techniques for evaluating both model fit and adequacy continue to provide fruitful areas for methodological research. However, these tools are best viewed as heuristic guides, rather than statutes. As such, it is important to use these measures thoughtfully and selectively, with attention to their limitations.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Academiai Kiado.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Berkhof, J., & Snijders, T. A. B. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics, 26*(2), 133–152. https://doi.org/10.3102/10769986026002133

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–270. https://doi.org/10.1007/BF02294361

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods &amp; Research, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecological Sociobiology, 65*, 23–35. https://doi.org/10.1007/s00265-010-1029-6

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.

Dominicus, A., Skrondal, A., Gjessing, H. K., Pedersen, N. L., & Palmgren, J. (2006). Likelihood ratio tests in behavioral genetics: Problems and solutions. *Behavior Genetics, 36*(2), 331–340. https://doi.org/10.1007/s10519-005-9034-7

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology, 44*(1), 205–231. https://doi.org/10.1006/jmps.1999.1284

Goldstein, H. (2010). *Multilevel statistical models* (4th ed.). Wiley.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B (methodological), 41*(2), 190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x

Hurvich, C. M., & Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science+Business Media.

Lee, H., & Ghosh, S. K. (2009). Performance of information criteria for spatial models. *Journal of Statistical Computation and Simulation, 79*(1), 93–106. https://doi.org/10.1080/00949650701611143

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 245–272). Information Age.

McCoach, D. B., & Cintron, D. W. (2021). *Introduction to modern modelling methods.* Sage.

McCoach, D. B., Newton, S. D., & Gambino, A. (2022). Evaluating the fit and adequacy of multilevel models. In A. A. O'Connell, D. B. McCoach, & B. A. Bell (Eds.), *Multilevel modeling methods with introductory and advanced applications.* Information Age Press.

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A., & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics, 43*(5), 594–627. https://doi.org/10.3102/1076998618776348

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics, 5*(4)*,* 746–762. https://www.jstor.org/stable/2958794

Palardy, G. J. (2011). Review of HLM 7. *Social Science Computer Review, 29*(4), 515–520. https://doi.org/10.1177/0894439311413437

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Raudenbush, S., Bryk, A., Cheong, Y., & Congdon, R. (2000). *HLM manual.* SSI International.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods, 24*(3), 309–338. https://doi.org/10.1037/met000018

SAS Institute Inc. (2020, December 23). *SAS help center: The MIXED procedure.* https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_mixed_toc.htm&docsetVersion=15.1&locale=en

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2)*,* 461–464. https://www.jstor.org/stable/2958889

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. https://doi.org/10.1007/BF02294360

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association, 82*(398), 605–610. https://doi.org/10.2307/2289471

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Chapman & Hall/CRC Press.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11*(4), 439–455. https://doi.org/10.1037/1082-989X.11.4.439

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50*(4), 1171–1177. https://www.jstor.org/stable/2533455

Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrerro, E. (2011). Selecting the best unbalanced repeated measures model. *Behavior Research Methods, 43*(1), 18–36. https://doi.org/10.3758/s13428-010-0040-1

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin &amp; Review, 11*, 192–196. https://doi.org/10.3758/BF03206482

Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The performance of IRT model selection methods with mixed-format tests. *Applied Psychological Measurement, 36*(3), 159–180. https://doi.org/10.1177/0146621612440305

Whittaker, T. A., Chang, W., & Dodd, B. G. (2013). The impact of varied discrimination parameters on mixed-format item response theory model selection. *Educational Psychological Measurement, 73*(3), 471–490. https://doi.org/10.1177/0013164412472188

Whittaker, T. A., & Stapleton, L. M. (2006). The performance of cross-validation indices used to select among competing covariance structure models under multivariate nonnormality conditions. *Multivariate Behavioral Research, 41*(3), 295–335. https://doi.org/10.1207/s15327906mbr4103_3

Zucchini, W., Claeskens, G., & Nguefack-Tsague, G. (2011). Model selection. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 830–833). Springer. https://www.encyclopediaofmath.org/images/b/b4/Model_selection.pdf

D. Betsy McCoach, Ph.D. is a Professor of Research Methods, Measurement, and Evaluation in the Educational Psychology department at the University of Connecticut. She teaches graduate courses in Structural Equation Modeling, Multilevel Modeling, Advances in Latent Variable Modeling, and Instrument Design. Dr. McCoach has co-authored over 100 peer-reviewed journal articles, book chapters, and books, including Instrument Design in the Affective Domain and Introduction to Modern Modeling Methods. Dr. McCoach is the founder and Program Chair of the Modern Modeling Methods conference, held at UCONN. Dr. McCoach's research interests include latent variable modeling, multilevel modeling, longitudinal modeling, instrument design, and gifted education.

Sarah D. Newton, Ph.D. is a Postdoctoral Research Associate in the University of Connecticut's Department of Educational Psychology. She completed her Ph.D. and M.A. in Educational Psychology (concentration in Research Methods, Measurement, & Evaluation) from the same institution. Sarah also earned an M.S. in Criminal Justice and a B.A. in Criminology, with completed course requirements in Psychology, from Central Connecticut State University. Her research focuses include model/data fit and model adequacy as complementary tools for multilevel model evaluation and selection; information criteria performance in multilevel modeling contexts; latent variable modeling; instrument design; reliability and validity theory; and quantitative research methodology.

Anthony J. Gambino, Ph.D. is a Postdoctoral Research Associate in the University of Connecticut's Department of Educational Psychology. He completed his Ph.D. and M.A. in Educational Psychology (concentration in Research Methods, Measurement, & Evaluation) from the same institution. His areas of expertise/research include causal inference, research and evaluation methodology, multilevel and structural equation modeling, and measurement theory.

Chapter 4
Concepts and Applications of Multivariate Multilevel (MVML) Analysis and Multilevel Structural Equation Modeling (MLSEM)

Yang Yang, Mengchen Su, and Ren Liu

Abstract Multilevel or hierarchical models (MLM, HLM) are widely used in analyzing the nested data in educational research over the past decades. The current trend in the research includes systematical thinking of the relationships in education (e.g., ecosystem model of human development) and continuous measurement on individual's performance (e.g., formative math assessment). These research focuses require that traditional analyses—structural equation modeling and multivariate analysis, work together with the MLM/HLM. This chapter will introduce two state-of-the-art techniques, which refer to multivariate multilevel (MVML) analysis and multilevel structural equation modeling (MLSEM). First, we will present the concepts of the MVML and MLSEM, emphasizing the ideas rather than algebra, to establish a theoretical and methodological foundation of the models. The MVML analysis allows for an analysis of multiple outcomes simultaneously, which can decompose the residual variances and covariances among outcome indicators into different levels. The MLSEM provides more appropriate estimates compared with SEM by considering the intra-class correlations. Then, we will discuss how the models can be applied to educational research. We expect both readers and researchers can find a more extended coverage of basic and advanced modeling techniques and get an enriched understanding of the value of MLM for future studies.

Keywords Nested data · Multivariate multilevel analysis · Multilevel structural equation modeling · Hierarchical data · Multilevel data

4.1 Introduction of Multilevel Modeling

Multilevel or Hierarchical structure data is common in the education sector (Woltman et al., 2012). For example, in school effects studies, researchers may be interested

Y. Yang (✉)
Beijing Normal University, Zhuhai, China
e-mail: yyang@bnu.edu.cn

M. Su · R. Liu
University at Buffalo, Buffalo, NY, USA

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_4

in how school-related factors, such as school characteristics (e.g., location, school composition, school's average socioeconomic status, etc.), school climate, or school educational resources, affect individuals' learning performance. Variables are measured and analyzed at both student and school levels. Additionally, multilevel structure data are used to study student development, where repeated measures/variables are collected for a group of persons at different time points, such as the studies investigating students' growth trajectory of reading performance during middle school. Observations with repeated measures/variables are nested within persons. The basic assumption of independence of observations is violated because students are not truly independent within the same classrooms or schools. Thus, researchers realize not only the importance of the hierarchical structure data, but also pay attention to what statistical techniques should be used to analyze such structured data without bias. Taking into account the unique structural feature of multilevel data, multilevel/hierarchical models (MLM/HLM) are introduced and developed in the early twentieth century (Hall & Malmberg, 2020). Unlike the single-level linear regression model, the MLM/HLM allows to investigate variance among variables at different levels simultaneously.

Although MLM has advantages over the OLS regression while analyzing nested data, conventional MLM techniques have limitations when dealing with complex relationships among independent and dependent variables, such as mediation effects among multiple related constructs, the relationship between latent constructs measured by a set of indicators, and the studies involved more than one outcome variables (Lee et al., 2018). In addition, the conventional MLM will produce bias estimates of between-cluster effects while analyzing within- and between-cluster effects (Lee, et al., 2018). Thus, for overcoming the limitations of MLM techniques, the multilevel models have combined with other statistical techniques, such as structural equation modeling (SEM) and multivariate analysis, and had widely implemented in educational research. The potential contribution of Multilevel Structural Equation Modeling (MLSEM) and Multivariate Multilevel Models (MVML)to educational research was recognized early in the development of the models. In the following sections, we will introduce the two advanced MLM/HLM techniques: MLSEM and MVML with examples from literature to help potential audience get a sense of how MLM/HLM could be worked together with SEM and Multivariate Analysis.

4.2 Multilevel Structural Equation Modeling (MLSEM)

Multilevel structural equation modeling (MLSEM) unites two statistical techniques, namely, Multilevel (or Hierarchal Linear) Modeling (MLM/HLM) and Structural Equation Modeling (SEM), which are common in educational research journals today. Since the first application of MLSEM in educational research was encountered via… *a multilevel path analysis on educational data in the Netherlands*, the number of studies using MLSEM in educational research increases steadily in the past three decades (Hall & Malmberg, 2020). MLSEM extends SEM by novel application to the nested data (e.g., students from the same class), which is typical in educational research mentioned above. Thus, MLSEM could be considered as the approach to handle nested data within an SEM in terms of, 1) adjusting standard errors of statistical estimates by taking the dependence of sampling into account, and 2) characterizing one model at each level of analysis (e.g., student-level model and classroom-level model). The former can be regarded as a simple modification of SEM to make estimates of statistical probability robust to dependence. Though fast to implement, this "flat" SEM approach lacks the ability to answer research questions regarding measurement and relationships that may vary across levels (e.g., individual students versus whole classes) with the same level of detail as the MLSEM alternative. For example, in answering research questions regarding whether the relationship between student academic achievement and instruction strategies is consistent across classrooms, or the extent to which the mechanism of an educational intervention holds true across different schools.

The application of MLSEM in educational research is twofold. First, MLSEM has a unique application to research questions relating to variation in measurement and classification across groups and/or time. Where there are unobserved constructs to be measured (e.g., via one of the many types of factor analyses, IRT, Rasch modeling) or latent groups to be identified then MLSEM provides a statistical framework to explore the extent to which that "latent" measurement and/or "latent" group identification is consistent over groups and/or time (Hall & Malmberg, 2020). For example, does the "latent" psychological construct of student self-efficacy remain conceptually equivalent over time and/or across students in different countries? Alternatively, to what extent is the composition of a group of students identified as "disadvantaged" consistent across different schools? Second, MLSEM has a unique application to a number of research questions relating to models in which a concept can both be dependent and independent on other, for example to research questions postulating mediation effects and/or indirect effects (e.g., Preacher et al., 2010). Again, when such questions consider variation across groups and/or time then MLSEM can have unique applications that are unavailable to other statistical approaches. Moreover, such unique applications of MLSEM can be combined with those mentioned in the paragraph above. This is illustrated by extending the example research questions just considered. For example, does student self-efficacy mediate the relationship between teacher effectiveness and student achievement and if so, how does this differ over time and/or across countries?

4.3 Examples of MLSEM Application

Example 1

The first example illustrated a simple application of MLSEM in exploring the extent to which a latent measurement was consistent in the student- and class-level estimation (1st application mentioned above). In the study of Bellens and colleagues (2019), one research question explored was the reliability of the measurement of instructional quality as classroom constructs in TIMSS (Trends in Mathematics and Science Study) 2015. To answer the research question, the authors used data from Flanders (5404 students from 295 classes), Germany (3948 students from 214 classes), and Norway (4164 student from 219 classes), and a two-level confirmatory factor analyses (CFA) model in Mplus 7.0 (Muthén & Muthén, 2012). A simplified mode could be found in Fig. 4.1.

The measurement of instruction quality contained three dimensions, namely, classroom management ([CM] 5 items), supportive climate ([SC] 10 items), and cognitive activation ([CA] 7 items). To evaluate the reliability of the constructs at student- and class-level simultaneously, the authors adopted a two-step analysis. First, intra-class correlation coefficients (ICCs) were calculated as an indication of the extent to which the dimensions of instruction quality can be seen as class-level constructs. Second, the CFA model (shown in Fig. 4.1) was estimated with omission of the missing data (Flanders 82 cases, Germany 635 cases, and Norway 28 cases were missing).

$$\mathrm{ICC}[1] = \frac{\tau^2}{\tau^2 + \sigma^2}$$

$$\mathrm{ICC}[2] = k \times \frac{k \times ICC[1]}{1 + (k-1) \times ICC[1]}$$

Where ICC [1] is the proportion of variance at the class level, τ^2 represents the variance between classes, σ^2 is the variance within classes (Raudenbush & Bryk, 2002), ICC [2] is evaluation of the reliability of the class mean, taking the class size (k) into account (Bliese, 1998).

The results of step one, take Norway for example (only Norway contained an analysis of all three dimensions), showed that ICCs [1] of CM, SC, and SA were

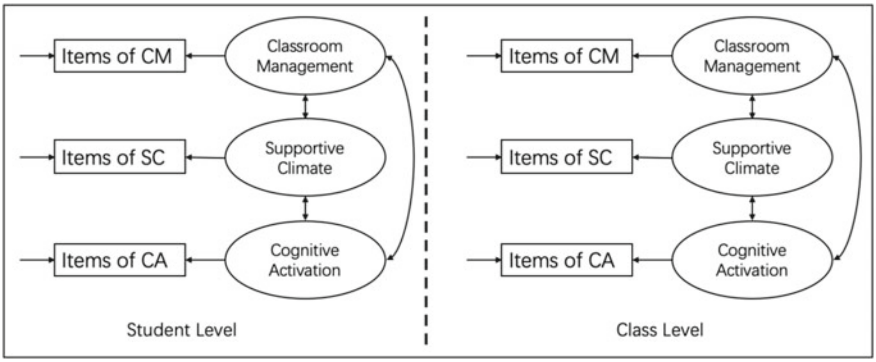


Fig. 4.1 Initial estimated CFA model to research questions one. *Source* Modified from the original figure

0.24, 0.14, and 0.05 respectively, which meant the proportion of variance at the class level for CM, SC, and SA was 24, 14, and 5%, respectively. The accordingly ICCs [2] were calculated as 0.86 for CM, 0.75 for SC, and 0.48 for CA. According to Schneider et al. (1998), an ICC [2] > 0.60 indicates a group-level construct. Therefore, the used items reliably captured the underlying constructs of CM and SC as class-level constructs in the educational system in Norway, while the dimension of CA needed further modification to become a reliable class-level construct. For the results of step two, although the model fit of significant indicators in the original model, including RMSEA, CFI, SRMR within, SRMR between, and Chi-square (df), was acceptable, the three modified models tested illustrated that the best model occurred when removing some items of low factor loadings and excluding the dimension CA from the within and between level.

Based on the results, the measurement of CM and SC was reliable as a class-level construct when the data was collected from individual students. But CA was problematic in distinguishing the differences between teachers' cognitive activation from the perspective of students. The segment of this study sheds light on measurement in educational research and researchers should be aware of the reliability of measurement when group effect existed.

Example 2

Different from the first example, in which the MLSEM was more data-driven, the second example involved models developed based on theory and previous research (2nd application mentioned above). In this study, Yang and colleagues (2020) examined the relationships between science teachers' professional development (PD) and their students' understanding of science concepts, with consideration of other factors and paths in theories and previous studies. The theoretical model is shown in Fig. 4.2. The model illustrates the possible theoretical path of how PD of teachers might eventually benefit students' learning with the presence of school environment, teacher and student characteristics (Desimone, 2009).

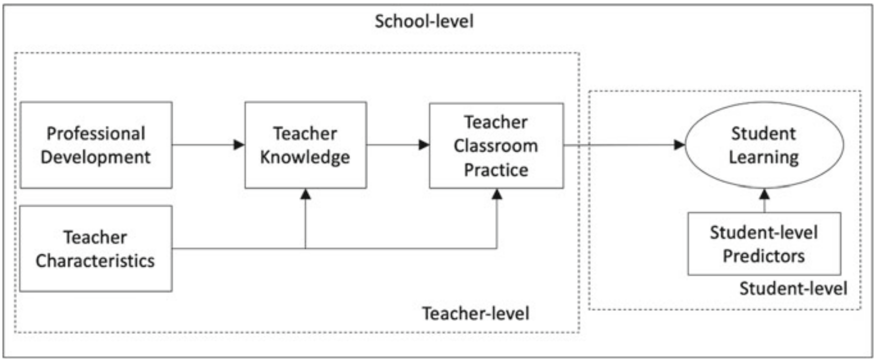


Fig. 4.2 Theoretical framework of professional development

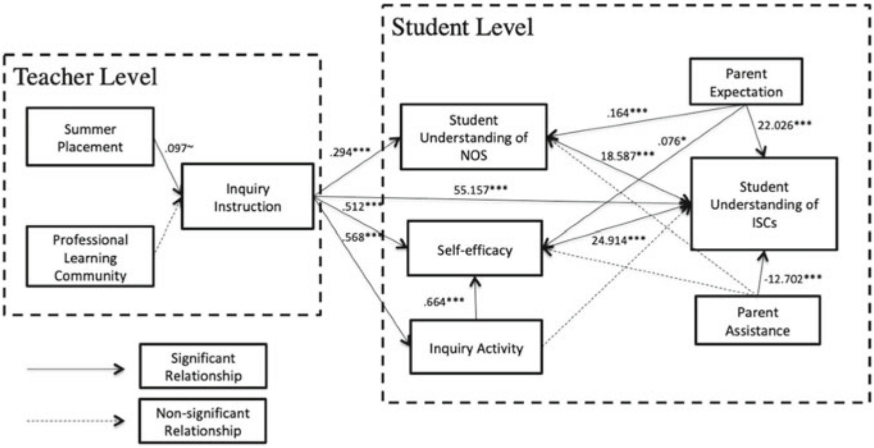


Fig. 4.3 Results of two-level path analysis

The data was collected from 200 teachers and their 5500 students within 12 local public schools. By considering the results of a pilot study of Yang and colleagues (2018) and an exploratory analysis of a three-level hierarchical linear model (HLM) according to the theory, the school-level factors were removed from the tested model because of between-school variance was negligible. Thus, a two-level MLSEM with ordinary least square (OLS) approach was adopted by using Mplus 7.0 software (Muthén &Muthén, 2012). The results of the model are shown in Fig. 4.3.

The model fit of the two-level path model was acceptable, with RMSEA = 0.028, CFI = 0.982, TLI = 0.953, SRMR within = 0.024, and SRMR between = 0.068, considering the standards of a good fit was that CFI and TLI > 0.900, RMSEA/SRMR < 0.800 (Van de Schoot et al., 2012). The path model generally verified the theoretical model and illustrated some detailed relationships. For example, even well-organized PD was not necessarily related to teacher classroom practice in terms of inquiry instruction, or in other words, the influence of PD on teacher instruction was more specific than expected. In this example, teachers' summer placement, which meant teacher joined research project and work with university science and engineering faculty for 6–8 weeks in summer vocation, slightly associated with teachers' practice of science inquiry, while teachers' monthly gathering in professional learning community to share experience in inquiry instruction, join instructional activities leading by university faculty, and so forth might not relate to teachers' practice in classroom in terms of scientific inquiry.

The tested model (Fig. 4.3) also illustrated some mediation effects of student-level factors, such as self-efficacy in science learning and experience in inquiry activity, considering the group effects of students in the same class.

4.4 Issues That Should Be Considered When Reporting MLSEM Applications

Considering the increasing use, complexity, and relative newness of MLSEM, researchers are required to take great care in the description of their methods and results. Issues of implementation of MLSEM could be found from a number of studies of SEM (e.g., MacCallum & Austin, 2000) or MLSEM (Schreiber & Griffin, 2004), and manuals, e.g., *Publication manual of the American Psychological Association* (American Psychological Association, 2020) and *Hierarchical linear models: Applications and data analysis methods* (Raudenbush & Bryk, 2002). Dedrick and colleagues (2009) proposed concise guidelines for reporting MLSEM applications: (1) Provide a clear description of the process of how the models were developed, such as how the predictors were selected and how many models were examined; (2) explicitly state the centering method of the variables; (3) explicitly state whether distributional assumptions were considered and how whether data were screened for outliers; (4) state the completeness of the data and how the missing data, if any, was treated; (5) provide detailed description of analysis methods, such as software (version) and method of estimation; (6) provide a complete list of estimations of all parameters; (7) provide either standard errors or confidence intervals (CIs) for the parameters of interest (Dedrick et al., 2009).

4.5 Multivariate Multilevel Model (MVML)

Working on the data in a nested (hierarchical) structure, sometimes, a researcher may be interested in more than one dependent variable, which is devoted to multivariate multilevel modeling. For example, if a study wants to examine students' achievement in various subjects simultaneously that is clustered within individuals (level-2) and groups (level-3), in this case, the multivariate multilevel model (MVML) will be conceptualized as a three-level model with the measurement model at the first level (Hox, 2002; Snijders & Bosker, 2012). This section introduces the basic concept and assumptions of MVML. Subsequently, the specification and application of the random intercept model are discussed with examples of educational studies (Kiwanuka et al., 2017; Lee et al., 2021; Su & Lee, 2021).

4.6 Basic Concepts and Assumptions of the Multivariate Multilevel Model

Multivariate multilevel model is the multilevel model that is one type of multilevel regression model with multiple outcome variables, which can be multiple facets or subscales of a construct or repeated interrelated measurements within individuals.

So why should we choose a more complicated multivariate multilevel model over a series of univariate models? Previous studies and books indicated MVML has several advantages over the univariate models (Hox, 2002; Hox et al., 2017; Snijders & Bosker, 2012; Tabachnick et al., 2007):

1. *More statistical power*: MVML has more power than univariate analysis. Hox and his colleagues (2017) suggested that the combined effect of a set of measurements can be statistically significant; however, the effect may be insignificant as using a single response measurement. Similarly, Snijders and Bosker (2012) made the same argument. They further explained that the additional power is more considerable if the outcome measures "are strongly correlated while at the same time the data very incomplete" (p. 283), which can be seen from the smaller standard errors in MVML.
2. *Less Type I error*: MVML has a lower rate of Type I error (rejecting a null hypothesis when it is true) than a series of univariate models.
3. *To test the joint effect*: if a study aims to test the joint effect of several facets of a construct or measurement simultaneously, it is necessary to use the MVML for reducing the risk of capitalizing on chance.
4. *Intercorrelations among outcomes*: MVML considers intercorrelations among multiple outcome variables. Hox and his colleagues (2017) indicated that "it is possible (for MVML) to test the equality of regression coefficients or variance components by imposing equality constraints" (p. 188).

As having many advantages over the univariate analysis, MVML has some disadvantages. First, MVML is far more complicated than separated analyses on each outcome variable. Although using MVML considers the interrelationship among multiple outcomes and has more power than univariate analysis, the model does not produce the indirect effect in the MSEM, which has a greater complexity for interpretation. Also, developing MVML into a more sophisticated model may have disadvantages compared with other advanced multilevel models, such as the multilevel structural equation model (MSEM). An example of the study by Su (2021) applied the MVML with a student-level mediator for testing the multilevel mediated relationship between school climate and multiple student outcomes through student engagement. Thus, before processing with MVML, a researcher should evaluate its pros and cons compared with other models and strategies.

4.7 Multivariate Random Intercept Model and Examples

The MVML is a three-level model with the measurement level at level-1. Suppose we have $p = 1\ldots, t$ measures nested among $i = 1\ldots$, $\mathrm{n_j}$ students within $j = 1\ldots$, J schools. Totally, we have p cases. indicate the multiple different dependent variables, we need to create dummy variables ($\mathrm{d_1}\ldots$, $\mathrm{d_t}$) at level-1 (measurement) model. $\mathrm{d_{ptij}}$ will be 1 or 0 depending on whether the response indicates a certain dependent variable Yor others, which is expressed as:

$$\mathrm{d}_{ptij} = \begin{cases} 1, (p = t) \\ 0, (p \neq t) \end{cases}$$

With p dummy variables, the level-1 (measurement) model must exclude the intercept term since it is not meaningful, which can be represented as:

$$Y_{tij} = \pi_{1ij} d_{1ij} + \pi_{2ij} d_{2ij} + \ldots \pi_{tij} d_{tij}$$

Here, it is noticeable that there has no error term in the measurement-level equation. Based on the level-1 (measurement) model, individual-level indicators are represented by $\beta_{t0j} \ldots \beta_{tij}$. The level-2 (individual) model is shown by:

$$\pi_{tij} = \beta_{t0j} + \beta_{t1j} + \ldots \beta_{tij} + \delta_{tij}$$

Group-level indicators are represented by γ_{tij}. For the random intercept model (all slope fixed) the level-3 (individual) model is shown by:

$$\beta_{t0j} = \gamma_{t00} + \gamma_{t01} + \ldots\gamma_{t0j} + \mu_{t0j}$$
$$\beta_{t1j} = \gamma_{t10}$$
$$\ldots$$
$$\beta_{tkj} = \gamma_{tk0}$$

4.8 Examples of MVML Applications

Example 1: Demographic Background, Academic and Socioeconomic Learning

Under Every Student Succeeds Act (ESSA), the new school accountability expands from focusing only on academic development to incorporating non-academic outcomes. The study by Lee et al. (2021) builds on the premise that academic and non-academic competencies are both important for reassessing K-12 school effectiveness and accountability. The study builds the multivariate multilevel model based on the Early Childhood Longitudinal Study-Kindergarten (ECLS-K): 2011 dataset. The multi-objective valued-added measures (MOVAM) are used to evaluate school effectiveness through academic and socioemotional domains. The study indicates MOVAM “applies a multivariate and multilevel model of school effectiveness to optimize the achievement of multiple learning objectives in a multi-layered school system” (Lee et al., 2021, p. 2).

The analytical sample contains 8957 young children in 771 schools in the Early Childhood Longitudinal Study (ECLS-K): 2011 data. Outcome variables are measured at the final round in Spring term, Grade three. The study has multiple outcomes variables, which include three academic skills (math, reading, and science), and three socioemotional skills (approach to learning, interpersonal ability, self-control). Student-level indicators include the measures of gender, race/ethnicity, English Language Learner (ELL) status, special education (SPED) status, and poverty (measured by whether students received free or reduced-price lunch). School-level aggregate variables aim to capture the compositional (group) effects beyond the student-level demographic variables.

Since the study examines multiple outcome variables simultaneously, MVML is constructed with the measurement level as level-1, student level as level-2, and school level as the level-3. Hierarchical Linear Modeling (HLM) 6.08 program are used to conduct the multivariate multilevel analysis. The conceptualized three-level model is shown as below.

Level-1 (measurement level)

Lee et al. (2021) describe that Y_{tij} and π_{tij} are "the observed and true score of student i in school j" on subject m. It is noticeable that the study considers the difference reliability estimates of the six outcome measures ($p = t = 1\ldots, 6$). To address the differences in reliability estimates (r), both sides of the formula need to be divided by its standards errors of measurement (SEM), which can be computed as follows: $\text{SEM} = \text{SD}\sqrt{(1-r)}$ (Snijders & Bosker, 2012). The variance at the level-1 needs to be fixed to 1 instead of 0.

$$Y_{tij} = \sum_{p=1}^{6} d_{ptij}\pi_{pij} + e_{tij}$$

d_{ptij} will be 1 or 0 depending on whether the response indicates a certain dependent variable Y or others, which is expressed as:$d_{ptij} = \begin{cases} 1, (p = m) \\ 0, (p \neq m) \end{cases}$

Level-2 (student-level)

$$\begin{aligned} \pi_{tij} =& \beta_{t0j} + \beta_{t1j}(pretest)_{tij} + \beta_{t2j}(Gender)_{tij} \\ &+ \beta_{t3j}(Race)_{tij} + \beta_{t4j}(FRPL)_{tij} + \beta_{t5j}(ELL)_{tij} \\ &+ \beta_{t6j}(SPED)_{tij} + \delta_{tij}(t = 1\ldots, 6) \end{aligned}$$

where

π_{tij} is the outcome t for student i in school j;
(pretest)$_{tij}$ is pretest score of student i in school j on subject t;
(Gender)$_{tij}$ is the "gender" gap (1 = female, 0 = male) of student i in school j on outcome t;
(Race)$_{tij}$ is the "minority" gap (1 = minority, 0 = non-minority) of student i in school j on outcome t;
(FRPL)$_{tij}$ is the "poverty" gap (1 = received free or reduced-price lunch, 0 = not eligible) of student i in school j on outcome t;
(ELL)$_{tij}$ the "language minority status" gap (1 = English language learner, 0 = not English language learner) of student i in school j on outcome t;

$(\mathrm{SPED})_{tij}$ the "special education status" gap (1 = special education (SPED), 0 = not SPED) of student *i* in school *j* on outcome *t*;

Level-3 (school-level)

$$\beta_{t0j} = \gamma_{t00} + \gamma_{t01}(Pretest_sch)_{tj} + \gamma_{t02}(\%Female)_{tj} + \gamma_{t03}(\%Minority)_{tj} + \gamma_{t04}(\%Poverty)_{tj} + \gamma_{t05}(\%Disability)_{tj} + \gamma_{t06}(\%ELL)_{tj} + \mu_{t0j}$$

$$\beta_{t1j} = \gamma_{t10}, \cdots, \beta_{tkj} = \gamma_{tk0}\ (k = 2\ldots6)$$

where

$(\mathrm{Pretest}_s\mathrm{ch})_{tj}$ is aggregated pretest score in school *j* on outcome *t*;
$(\%\mathrm{Female})_{tj}$is the percentage of female students in school *j* on outcome *t*;
$(\%\mathrm{Minority})_{tj}$is the percentage of minority students in school *j* on outcome *t*;
$(\%\mathrm{Poverty})_{tj}$ is the percentage of students received reduced/free lunch in school *j* on outcome *t*;
$(\%\mathrm{Disability})_{tj}$ the percentage of students in special education in school *j* on outcome *t*;
$(\%\mathrm{ELL})_{tj}$ the percentage of students who are English language learners in school *j* on outcome *t*;

To be a multivariate random intercept model, all slopes at level-3 (school-level) need to be fixed, and μ_{t0j} "is assumed to vary across schools around grand mean after holding constant the school-level pretest score and demographic covariates" (Lee et al., 2021, p. 4).

The reprinted Table 4.1 presents the fully conditional model (with student- and school-level predictor) results reported by Lee et al. (2021). The table provides information on both fixed and random effects. The results of fixed effects show different student-level patterns across six dependent variables. For example, there was a significant difference in students' average reading achievement between female and male students. On average, female students' math achievement score was 1.44 points lower than males (β_{12} =1.44, s.e. = 0.24, $p < 0.001$). The interpretation for continuous predictors is different. Students' mean pretest score is a statistically significant and positive predictor for their reading achievement. On average, one-unit increase in students' pretest score predicted 0.30 points higher score in students' reading achievement (β_{11} =0.30, s.e. = 0.01, $p < 0.001$). Regarding the school organizational features, for example, the percentage of female students in school positively predicts students' interpersonal skills. A one-unit increase in the school mean of the percentage of female students is associated with 0.18 points increase in students' interpersonal skills (β_{6i2} =0.18, s.e. = 0.08, $p < 0.05$). The results also show how much variance in each dependent variable are explained by student-level and school-level predictors. For example, the results suggest that there are less than/around 50% of variance in academic outcomes are explained by student-level predictors (33.58–57.84%). School-level predictors explained more than two-thirds of the variances (62.72–76.17%).

Table 4.1 Estimated student and school-level effects on academic and socioemotional outcomes

	Reading	Math	Science	Approach to learning	Self-control	Interpersonal skills
Fixed effects						
Level-2 (student level)						
Pretest	0.30*** (0.01)	0.47*** (0.01)	0.59*** (0.01)	0.25*** (0.01)	0.22*** (0.01)	0.20*** (0.01)
Female	1.44*** (0.24)	− 2.94*** (0.26)	− 1.34*** (0.20)	0.28*** (0.02)	0.20*** (0.02)	0.27*** (0.02)
Minority	− 0.64 (0.38)	− 2.74*** (0.40)	− 1.13*** (0.27)	− 0.01 (0.03)	0.00 (0.02)	0.00 (0.03)
Poverty	− 3.76*** (0.36)	− 3.36*** (0.37)	− 2.39*** (0.27)	− 0.18*** (0.03)	− 0.15*** (0.02)	− 0.15*** (0.02)
Disability	− 1.32*** (0.67)	− 9.45*** (0.86)	− 4.48*** (0.55)	− 0.31*** (0.04)	− 0.14*** (0.04)	− 0.18*** (0.04)
ELL	− 2.31*** (0.44)	0.65 (0.56)	0.14 (0.38)	0.11*** (0.03)	0.14*** (0.03)	0.08** (0.03)
Level-3 (school level)						
Pretest	0.27*** (0.03)	0.34*** (0.04)	0.61*** (0.04)	0.14*** (0.03)	0.15*** (0.03)	0.14*** (0.03)
%Female	3.11** (1.14)	− 1.12 (1.35)	− 0.18 (0.83)	0.23** (0.08)	0.19* (0.07)	0.18* (0.08)
%Minority	− 2.07** (0.78)	− 5.36*** (0.89)	− 0.89 (0.61)	− 0.04 (0.05)	− 0.11* (0.05)	− 0.05 (0.05)
%Poverty	− 7.75*** (0.84)	− 8.12*** (1.02)	− 6.07*** (0.69)	− 0.25*** (0.05)	− 0.23*** (0.05)	− 0.20*** (0.05)
%Disability	− 8.81** (2.84)	− 8.04* (3.18)	− 5.09* (1.95)	− 0.44** (0.14)	− 0.33* (0.13)	− 0.36* (0.14)
%ELL	− 1.27 (0.95)	4.01*** (1.17)	1.63* (0.74)	0.13* (0.06)	0.26*** (0.06)	0.16** (0.06)
Random Effects						
Level-2 variance	69.89***	76.59***	21.24***	0.31***	0.22***	0.28***
% Variance explained by level-2	33.58	46.07	57.84	26.28	20.31	18.12
Level-3 variance	7.43***	12.95***	5.02***	0.03***	0.03***	0.04***

(continued)

Table 4.1 (continued)

	Reading	Math	Science	Approach to learning	Self-control	Interpersonal skills
% Variance explained by level-3	69.06	62.72	76.17	6.77	18.54	9.42

Note *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$ Reprinted from Lee, J., Kim, T., & Su, M. (2021). Reassessing school effectiveness: Multi-objective value-added measures (MOVAM) of academic and socioemotional learning. *Studies in Educational Evaluation, 68,* 6

The study by Lee et al. (2021) is a well-designed example for random intercept MVML, which investigates both student- and school-level predictors on multiple outcome indicators. MVML considers the interrelationship among academic and socioemotional outcomes, which has more power than pairs of univariate analyses on each outcome.

Example 2: Student and Classroom Characteristics and The Multifaceted Structure Of Attitude Toward Mathematics

Kiwanuka et al. (2017) design a study that explored the relationships between students, classroom characteristics, and the multifaced structure of students' attitudes toward mathematics. Using a two-stage random sampling design, 4,819 grade 7 students were selected within 78 classes. SAS 9.3 with the ML method is used for the multivariate multilevel analysis.

The study targets three facets of students' attitudes toward mathematics (ATM) as outcome variables. Student-level predictors include SES, gender, age, prior Math achievement (PMA), parents' beliefs/attitudes (PABELF), math self-confidence (PRCONF), the usefulness of math (PRUSE), and enjoyment of Math (PRENJOY). Classroom-level indicators include the composite measures of the class mean of all student-level indicators. It also adds indicators like classroom learning environment (CLEARN), classroom assessment (CLASSESS), classroom questioning (CLQUEST), classroom modeling (CLMODEL), math teacher beliefs/attitudes (MTBELF), and peer influence (PEER). Similar to the study by Lee et al. (2021), MVML is constructed with the measurement level as level-1, student level as level-2, but classroom level as the level-3, which is shown as below:

Level-1 (measurement level)

Y_{tij} and π_{tij} are the observed and true score of student *i* in class *j* on the m facet of attitudes toward mathematics. Different from the first example, this study does not correct its reliability estimates based on the standard error of measurement ($p = t = 1\ldots, 3$), which is a limitation for its measurement-level modeling.

$$Y_{tij} = \sum_{p=1}^{3} d_{ptij}\pi_{pij} + e_{tij}$$

$\mathrm{d_{ptij}}$ will be 1 or 0 depending on whether the response indicates a certain dependent variable Y or others, which is expressed as: $d_{ptij} = \begin{cases} 1,\ (p = m) \\ 0,\ (p \neq m) \end{cases}$

Level-2 (student-level)

$$\pi_{tij} = \beta_{t0j} + \sum_{z=1}^{8} \beta_{tzj} K_{tzj} + \delta_{tij} (t = 1\ldots, 3)$$

where

π_{tij} is the outcome *t* for student *i* in class *j*;

K_{tij}represents one of the eight student-level predictors in this study (k = 1…,8; SES, gender, age, PMA, PABELF, PRCONF, PRUSE, PRENJOY).

Level-3 (Classroom-level)

$$\beta_{t0j} = \gamma_{t00} + \sum_{m=1}^{8} \gamma_{t0m} K_{clsmj} + \mu_{t0j}$$

$$\beta_{t1j} = \gamma_{t10}, \cdots, \beta_{tmj} = \gamma_{tmo}\,(m = 1, \ldots, 13)$$

where

K_{clsmj} represents one of the thirteen classroom-level predictors in the study ($m =$ 1…,13; CLEARN, CLASSESS, CLQUEST, CLMODEL, MTBELF, PEER, CLSES, CLGIRLS, CLPMA, CLPABELF, CLPRCONF, CLPRUSE, and CLPRENJOY). For the multivariate random intercept model, level-3 (classroom-level) slopes are fixed.

The study builds five models. First, the unconditional model is built without any indicator (Model 0). There are four conditional multivariate multilevel models: the first model includes the student-level characteristics; the second model adds class processes indicators; the third model incorporates class composition predictors; and the final model (fully conditional model) includes class composition all student-level and classroom-level variables. The reprinted Table 4.2 is based on the model 2 results from the study by Kiwanuka et al. (2017), which is showed for the illustrative purpose.

The study provides deviance (−2 Res Log Likelihood) of all models, which provides information on testing and comparing the model fit. The difference in deviance between model 1 and model 2 is: Chi-square statistics $\mathcal{X}^2(15) = 105{,}059.6 - 105{,}221.3 = 161.7$, $p<0.001$. The results support that the model is significantly improved as adding indicators of classroom processes. Table 4.2 also indicates student-level indicators explain 4.8, 3.0, and 4.8% of the variance in students' math self-confidence, usefulness of math, and enjoyment of math, respectively. Classroom-level indicators explain 73.7, 51.9, and 62.6% of the variance in students' math self-confidence, the usefulness of math, and enjoyment of math.

For the illustrative perspective, we look at two classroom processes indicators (Classroom modeling, classroom assessment), the results show, on average, classroom assessment is a statistically significant and negative predictor for all three attitudinal indicators ($\beta_{CONF} = -1.27$, $\beta_{USE} = -1.52$, $\beta_{ENJ} = -1.22, p < 0.05$). Classroom modeling is a significant and positive predictor for all three attitudinal indicators (β_{CONF} =1.66, β_{USE} =3.68, β_{ENJ} =2.15, $p < 0.05$). It is noticeable that MVML allows to control students' prior math self-confidence, the usefulness of math, and enjoyment of math accordingly. We can see student-level models do not have to be identical across the three outcome variables. Unlike the study by Lee et al. (2021), Example 2 shows how MVML can capture a multifaceted construct.

Table 4.2 HLM parameter estimates for student- and classroom-level predictors (Model 2)

	Model 2: student characteristics + class process		
	CONF	USE	Enjoy
Intercept	60.40*(0.57)	75.95*(0.82)	62.51*(0.75)
Level-2 (student level)			
SES	0.15(0.26)	0.06(0.28)	0.41(0.30)
Gender	−0.17(0.53)	2.84*(0.58)	−0.46(0.61)
Age	−0.76*(0.27)	−0.28(0.29)	−0.54(0.31)
PMA	1.90*(0.27)	0.68(0.29)	1.91*(0.31)
PABELF	0.94*(0.28)	1.22*(0.31)	1.54*(0.32)
PRCONF	2.26*(0.24)	–	–
PRUSE	–	1.57*(0.28)	–
PRENJOY	–	–	1.97*(0.26)
Level-3 (classroom level)			
CLEARN	−0.28(0.43)	−0.55(0.70)	−1.10(0.60)
CLASSESS	−1.27*(0.41)	−1.52*(0.70)	−1.22*(0.59)
CLQUEST	1.21*(0.52)	0.36(0.86)	2.06*(0.73)
CLMODEL	1.66*(0.49)	3.68*(0.80)	2.15*(0.69)
MTBELF	−0.67(0.44)	−0.68(0.70)	−1.40*(0.60)
PEER	0.96(0.51)	1.23(0.81)	1.59*(0.70)
% Variance explained by level-2	4.8	3.0	4.8
% Variance explained by level-3	73.7	51.9	62.6
−2 log likelihood	105,059.6 − 105,221		

Note *$p < 0.05$; Reprinted from Kiwanuka, K. N., et al. (2017). How do student and classroom characteristics affect attitude toward mathematics? A multivariate multilevel analysis. *School Effectiveness and School Improvement, 28*(1), 11

4.9 Summary

In the past decades, studies using hierarchical/multilevel data in the education sector have increased dramatically. The current trend in the research includes systematical thinking of the relationships in education (e.g., ecosystem model of human development) and continuous measurement on individual's performance (e.g., formative math assessment). These research focuses require that traditional analyses—SEM and Multivariate Analysis, work together with the MLM/HLM. In this chapter, we cover the concepts and applications of the MLSEM and MVML techniques to facilitate researchers to select appropriate statistical techniques for dealing with nested data. This chapter provides researchers a bigger picture of multilevel data analysis techniques. The MLSEM techniques combine the HLM and SEM techniques for dealing with how latent variables vary across groups/ organizations. The MLMV techniques incorporate multivariate analysis with MLM/HLM, which focus on analyzing nested data with various outcomes variables. The utility of those two advanced statistical analysis methods is now accelerating in educational research. Applying MLSEM and MLMV techniques might require extra learning for dealing with technical and practical challenges, but we would encourage readers to consider those advanced techniques while doing research with nested data.

Recommendations for Using MLSEM and MVML techniques.

MLSEM and MVML techniques are advanced and complicated multilevel modeling strategies. It is necessary to know what types of research questions/data should be analyzed using MLSEM and MVML techniques. First, MLSEM could be used to test the variation of latent variables/measures across groups or over time for individuals. On the one hand, the results of MLSEM could help validate and modify existing measurement instruments. Specifically, the ICCs help identified potential group effects on the reliability of the measurement. On the other hand, while developing new measurement instruments, applying MLSEM helps researchers to think about the hierarchical structure of data.

Another common application of MLSEM is to help identify mediation/indirect relationships among variables across groups or time. The path analysis in MLSEM techniques is different from traditional SEM because it requires researchers to think about the hierarchical structure of variables.

In terms of MVML, it has more power than univariate multilevel analysis when the studies consisted of various dependent variables (e.g., students' math, reading, science, and other non-cognitive outcomes) together, because it considers the interrelations among multiples outcome measures. In addition to the studies involved more than one outcome measures, MVML techniques could be used in the study with multifaceted/ multidimensional construct, such as student engagement which has been considered through behavioral, cognitive, and emotional dimensions. MVML techniques will take into account the interrelationships among facets/dimensions within one latent construct, which increases the accuracy of the analysis.

References

American Psychological Association. (2020). *Publication manual of the APA* (7th ed.). Author.

Bellens, K., Van Damme, J., Van Den Noortgate, W., Wendt, H., & Nilsen, T. (2019). Instructional quality: Catalyst or pitfall in educational systems' aim for high achievement and equity? An answer based on multilevel SEM analyses of TIMSS 2015 data in Flanders (Belgium), Germany, and Norway. *Large-Scale Assessments in Education, 7*(1), 1–27.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods, 1*(4), 355–373.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69–102.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.

Hall, J., & Malmberg, L. (2020). The contribution of multilevel structural equation modelling to contemporary trends in educational research. *International Journal of Research &amp; Method in Education, 43*(4), 339–347. https://doi.org/10.1080/1743727X.2020.1796066

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Lawrence Erlbaum.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications.* Routledge.

Kiwanuka, H. N., Van Damme, J., Van Den Noortgate, W., Anumendem, D. N., Vanlaar, G., Reynolds, C., & Namusisi, S. (2017). How do student and classroom characteristics affect attitude toward mathematics? A multivariate multilevel analysis. *School Effectiveness and School Improvement, 28*(1), 1–21. https://doi.org/10.1080/09243453.2016.1201123

Lee, J., Kim, T., & Su, M. (2021). Reassessing school effectiveness: Multi-objective value-added measures (MOVAM) of academic and socioemotional learning. *Studies in Educational Evaluation, 68*, 100972. https://doi.org/10.1016/j.stueduc.2020.100972

Lee, J., Shapiro, V. B., Kim, B. K. E., & Yoo, J. P. (2018). Multilevel structural equation modeling for social work researchers: An introduction and application to healthy youth development. *Journal of the Society for Social Work and Research, 9*(4), 689–719. https://doi.org/10.1086/701526

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201–226.

Muthén, L. K., & Muthén, B. O. (2012). Mplus version 7.0 [computer software]. Retrieved from https://www.statmodel.com/index.shtml

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209–233.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Sage.

Schneider, B., White, S. S., & Paul, M. C. (1998). Linking service climate and corner perceptions of service quality: Test of a causal model. *Journal of Applied Psychology, 83*, 150–163.

Schreiber, J. B., & Griffin, B. W. (2004). Review of multilevel modeling and multilevel studies in The Journal of Educational Research (1992–2002). *The Journal of Educational Research, 98*, 24–33.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Su, M. (2021). *The mediation effect of student engagement on the relationship between school climate, socioemotional well-being, and academic achievement: A cross-cultural comparative study of China, Korea, Finland, and the U.S.* (Order No. 28645024). Available from ProQuest Dissertations & Theses Global. (2580988311). https://www.proquest.com/dissertations-theses/mediation-effect-student-engagement-on/docview/2580988311/se-2?accountid=14169

Su, M., & Lee, J. (2021, April). *Whole-child education: China, Korea, Finland, and U.S. Students' Academic Achievement and Socio-Emotional Well-Being*. Paper presented at the 2021 annual meeting of American Educational Research Association (AERA), virtual conference.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5, pp. 481–498). Pearson.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52–69.

Yang, Y., Liu, X., & Gardella, J. A., Jr. (2020). Effects of a professional development program on science teacher knowledge and practice, and student understanding of interdisciplinary science concepts. *Journal of Research in Science Teaching, 57*(7), 1028–1057.

Yang, Y., Liu, X., & Gardella, J. (2018). Impact of professional development on teacher knowledge, practice and student understanding of science in an interdisciplinary science and engineering partnership. *Journal of Science Teacher Education, 29*(4), 263–282.

Yang Yang, Ph.D. Assistant Professor in Center for Educational Science and Technology, Institute of Advanced Studies in Humanities and Social Sciences, and Research Institute of Science Education (RISE), Beijing Normal University, China. The research area includes curriculum design and teacher professional development in science education, measurement & evaluation, and quantitative research methods.

Mengchen Su, Ph.D. in Educational Psychology and Quantitative Methods at SUNY University at Buffalo, the United States. Research areas specialize in multilevel modeling in comparative and whole child education, focusing on educational equity and accountability.

Ren Liu, Ph.D. in Curriculum, Instruction, and Science of Learning at the University at Buffalo, the State University of New York. Research areas specialize in educational measurement, college teaching evaluation and improvement, and quantitative research methods.

Chapter 5
Data Visualization for Pattern Seeking in Multilevel Modeling

Chong Ho Alex Yu

Abstract It is a common practice for educational researchers to employ multilevel modeling to analyze archival data that were collected by multistage sampling (e.g. Program for International Student Assessment (PISA), Trends for International Math and Science Study (TIMSS), High School and Beyond (HSB), etc.). It is noteworthy that usually the sample size of this type of international and national studies is extremely large, and thus its ultra-high statistical power is associated with a high Type I error rate. Instead of counting on the *p*-value alone to make a dichotomous decision (to reject or not to reject the null hypothesis), it is advisable to utilize data visualization for pattern seeking. The objective of this chapter is to illustrate how various data visualization techniques can enable researchers to extract insight from data at each step of multilevel modeling. Specifically, this chapter illustrates techniques including linking and brushing, binning and median smoothing, and usage of a bubble plot, local filter, analysis of mean plot, residual plot, and many others.

Keywords Data visualization · PISA · TIMSS · HSB · Achievement · Graph builder

5.1 Introduction

It is a well-known fact that many data sets collected in social science settings do not meet the parametric assumption of independence, particularly in national or global data sets. For example, in nationwide educational research, simple random sampling at the individual level could yield a biased sample, because more students would be selected from highly populated cities, such as Los Angles and New York. As a remedy, researchers employ multistage sampling schemes, such as cluster sampling and stratified sampling, to collect representative data (Thompson, 2012). These techniques involve partitioning the accessible population into different segments, such

C. H. A. Yu (✉)
Azusa Pacific University, Azusa, CA, USA
e-mail: cyu@apu.edu

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_5

as state, county, school district, and school, so that every corner of the country is covered.

If some clusters (e.g. certain counties in California) are randomly chosen and every element in the subset is sampled, this method is known as cluster sampling. If all the strata (e.g. all fifty states in the US) are included for subsequent random sampling, then this method is called stratified sampling. No matter whether cluster sampling, stratified sampling, or a hybrid of both is used, it is obvious that the observations can no longer meet the assumption of independence required by many parametric tests, such as Ordinary Least Squares (OLS) regression analysis. To be more specific, all the students from a particular school are influenced by the same teachers, superintendent, school policies, school resources, and local culture. By the same token, all the schools that are affiliated with a particular school district also share a common source of influence. As a result, the *F*-statistic yielded from OLS regression analysis tend to be too large and the standard error estimates tend to be too small (Bauer & Curran, 2017).

Besides independence, there exist other potential violations of regression assumptions that pose a challenge to the validity of conclusions. OLS regression assumes that the predictor levels are fixed, meaning that the researcher exclusively selects the levels of the predictor (Bauer & Curran, 2017). A drug test is a typical example. In this case the medical researcher is only interested in comparing the effects of Drug A, B, and C on cancer patients. No other drugs or treatments are under consideration, meaning that the conclusion would not be generalized to the drugs that were not tested. However, this is not the case for data collected using cluster sampling, because the levels of predictors (e.g. county, school district, and school) are randomly selected from the population. However, the researcher would like to generalize the conclusion to all students, not just the students in certain counties, school districts, and schools. Further, the assumption of homoscedasticity might not be met when there is a wide variation among counties, school districts, and schools in terms of the primary independent variable. In a fixed effect model, it is implicitly assumed that one size can fit all; this assumption overlooks between-group differences among counties, school districts, and schools (Bauer & Curran, 2017).

To rectify this situation by taking into account the nested structure of the data, many social scientists use hierarchical linear modeling, also known as mixed modeling or multilevel modeling (Raudenbush & Bryk, 2002). On the one hand, this procedure is innovative for its robustness against relatedness among observations. On the other hand, it is still considered traditional because the dichotomous decision of whether or not to reject the null hypothesis is still based upon an arguably arbitrary cut-off point, namely, alpha $< =$ 0.05. The arbitrariness of this decision presents a threat to the validity of the conclusion. Usually the sample size of a national or international study utilizing multistage sampling is large, resulting in very high statistical power. Thus, even a trivial effect could be identified as a significant finding. Although multilevel modeling is indispensable for analyzing hierarchical data, it is the conviction of the author that it should be used in parallel with data visualization, which aims to reveal the pattern of the data rather than simply yielding a dichotomous conclusion.

In this chapter, dynamic data visualization techniques for analyzing nested data are illustrated using the data set of High School and Beyond (HSB) collected by the National Center for Educational Statistics (2021). It is important to point out that data visualization is a form of exploratory data analysis; as such, researchers can be greater benefited by dynamic data visualization, when compared with static graphing. In the latter the output is not linked with other objects and cannot be further manipulated. In contrast, when a graph is dynamic, any change in one graph can trigger a corresponding change in all other objects of the same data file. Further, an analyst can alter a dynamic graph, such as by inserting another variable into the graph. This feature is important because asking what-if questions in an exploratory fashion can lead to insightful knowledge discovery (Yu, 2014, 2017). SAS Institute is a major leader in dynamic data visualization. As such, in this chapter JMP Pro developed by SAS Institute (2021) is used for illustration. Although SAS/STAT offers many more powerful procedures for multilevel modeling, JMP is recommended for users who would like to obtain results quickly by focusing on the content, rather than the programming syntax. If your institution does not have a site license of JMP Pro, you can download a trial version from http://www.jmp.com/try or purchase a faculty/student version from https://onthehub.com. The sample data used in this chapter can be downloaded from https://creative-wisdom.com/pub/hsb.jmp.

5.2 Data Source

High School and Beyond (HSB) is a longitudinal study launched by the National Center for Educational Statistics (2021) in 1980. The objectives of this project include examining student trajectories after leaving high school into postsecondary education, the workforce, and beyond. Researchers can use the data to identify factors that might influence the students' educational and career outcomes after passing through the US educational system. For instance, the socioeconomic status of a student's family is said to be a crucial predictor of mathematic achievement, which is a valuable skill set in an information-oriented economy. Prior research found that both men and women who took college preparatory math coursework in high school had lower unemployment at midlife (Bosky, 2019). To explore the relationship between SES and math competency, a subset of HSB ($n = 7185$) containing relevant variables is utilized throughout the entire chapter. The variables in the data set are explained in Table 5.1.

Table 5.1 Code book of HSB

Variable name	Description	Measurement level
student ID	A unique identifier for the student	Nominal
student_mathach	Mathematics achievement score of the student	Continuous
student_ses	The student's SES is a composite variable, synthesizing from a set of variables from the base-year and first follow-up data, including father's occupation, father's education, mother's education, family income, and material possessions in the household (NCES, 2020). This variable is centered by the grand mean	Continuous
student_cses	The student SES centered by the local school mean, rather than the grand mean	Continuous
school_meanses	The mean of student SES of each school	Continuous
student_min	A dummy variable that indicates whether the student is a minority	Nominal
student_female	A dummy variable that indicates whether the student is a female	Nominal
school_ID	A unique identifier for the school	Nominal
school_size	The variable indicates the size of the school	Continuous
school_sector	A dummy variable that indicates whether the school is public or private	Nominal
school_disclim	The disciplinary climate of the school is a composite score, consisting of scoring for the perception of these statements: "There are clear rules about student behavior, "discipline is fair," "everyone knows what the school rules are," and "school rules are applied equally to all students." The score is centered by the grand mean (NCES, 2018)	Continuous

5.3 Preliminary Data Visualization

5.3.1 Profile Analysis by Linking and Brushing

There is no doubt that the hierarchical structure of this data set adds an extra layer of complexity to the analysis. Nevertheless, as is the case with other analytical strategies, an analyst should start with a preliminary analysis of the multivariate relationship between the dependent and the independent variables, putting aside school level. Although the primary research question is about whether there is a meaningful relationship between SES and math achievement, an analyst can gain insights by examining all available data in an exploratory mode. Exploratory data analysis necessitates dynamic data visualization. As mentioned before, dynamic graphing in JMP Pro can link different objects (e.g. histograms) together. The steps to create linked histograms of all relevant variables are:

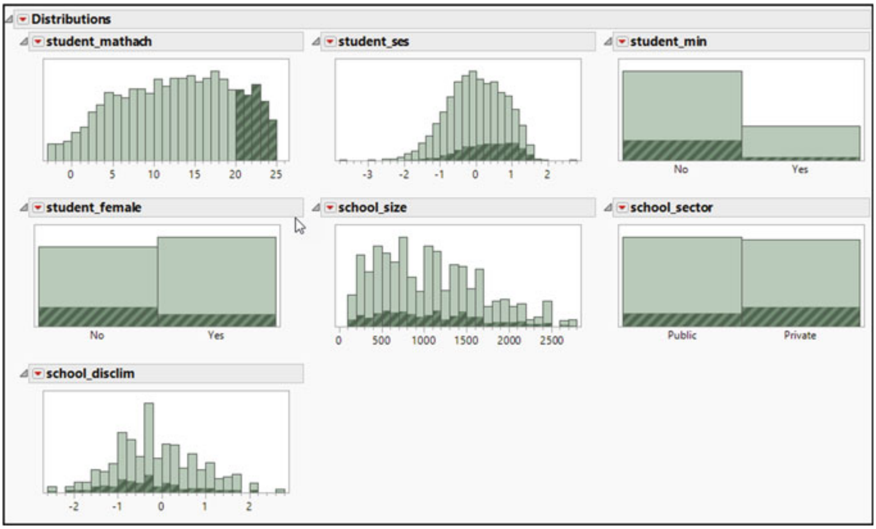


Fig. 5.1 Profile analysis of high math achievers by linking and brushing

1. Open **Distribution** from the pull-down menu **Analyze**.
2. Put student_mathach, student_ses, student_min, student_female, school_size, school_sector, and school_disclim into **Y, columns**.
3. Uncheck the box **Histograms only** (optional).
4. Press **OK**.
5. Hold down the shift key to select the students that scored 20 or above in student_mathach (brushing).

There is a shortcut in the data set: the user can click on the green arrow next to the script entitled "Distributions" located at the upper left hand corner of the data file. Figure 5.1 shows the distributions of all selected variables. The figure was altered so that all distributions can be fitted into this page; thus, its appearance is slightly different from the default. As the name of the technique implies, the "brushed" (selected) observations in one histogram are linked with the same observations in others. Figure 5.1 indicates the profile of high achievers in math tests: their SES ranges from −1 to +2, most of them are non-minorities, males, and private school students, but there is no clear pattern of the high achievers in terms of the school size and the school disciplinary climate. At first glance this procedure is not directly related to multilevel modeling. However, as illustrated in a later section, when a model fit evaluated by a plot of actual and predicted scores is less than ideal, an analyst can determine what other predictors should be included, in order to improve model fit.

5.3.2 Binning and Median Smoothing for Examining the Relationship between SES and Math Achievement

The primary research question is: Can SES predict math performance? A preliminary analysis can be conducted without taking into account the factor of school. At first glance, the logical choice for this task is regression analysis. However, with a sample of 7,185, it is not surprising to obtain a significant result ($p < 0.0001$). As a matter of fact, with such a large sample size, any effect—even if trivial—could be identified as significant. Table 5.2 shows the result of the multiple regression model, in which all available independent variables were used to predict math performance. Not surprisingly, all p-values are less than 0.05, but the results should be interpreted with caution. After investigating these so-called "significant" predictors, it was found that not all of them are meaningful. For ease of illustration this chapter focuses only on the relationship between SES and math achievement.

A common data visualization technique for examining the association between a dependent and an independent variable is the scatterplot (see Fig. 5.2). However, over-plotting—in which jammed data points appear to be a messy cluster of ink—obscures analysts from seeing any pattern (Yu & Behrens, 1995). In addition, it seems that most students hit a ceiling in terms of SES, and also there are several bivariate outliers. The remedy for these issues will be discussed in the section of using SES centered by the local mean. In this section let's focus on the problem of overplotting.

The binning approach suggested by Carr (1991) directly attacks the problem of overplotting. In this approach data points are grouped in intervals, thus reducing data congestion. Following this approach, one can search for a pattern by dividing the data into several portions along the x-dimension, computing the median of y in each portion, and then look at the trend by connecting the medians (Tukey, 1977). The steps for binning SES are as follows:

1. Select student_ses by clicking on the variable header.
2. Choose **Utilities/Make Binning Formula** from the pull-down menu **Cols**.
3. In the pop-up panel, the data are partitioned into several bins (intervals). However, one can see that there is only one observation in the first and last

Table 5.2 Results of regression analysis

Term	Estimate	Std error	t ratio	Prob > |t|
Intercept	11.236085	0.167971	66.89	< 0.0001*
student_min[No]	1.579636	0.085556	18.46	< 0.0001*
student_female[No]	0.7522634	0.073398	10.25	< 0.0001*
student_ses	2.2999083	0.099387	23.14	< 0.0001*
school_size	0.0007252	0.000134	5.43	< 0.0001*
school_sector[Public]	−0.871478	0.11057	−7.88	< 0.0001*
school_disclim	−0.66522	0.112297	−5.92	< 0.0001*

* Significant at the 0.01 level

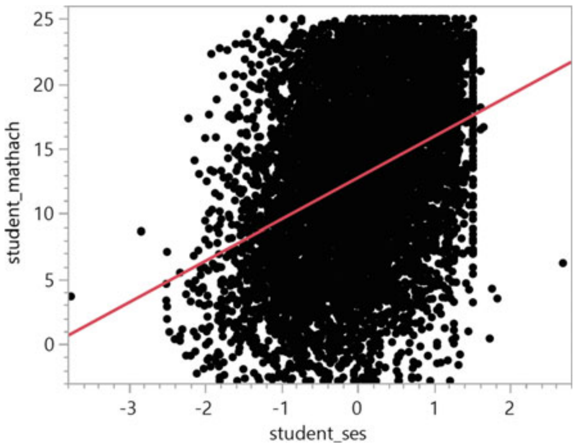
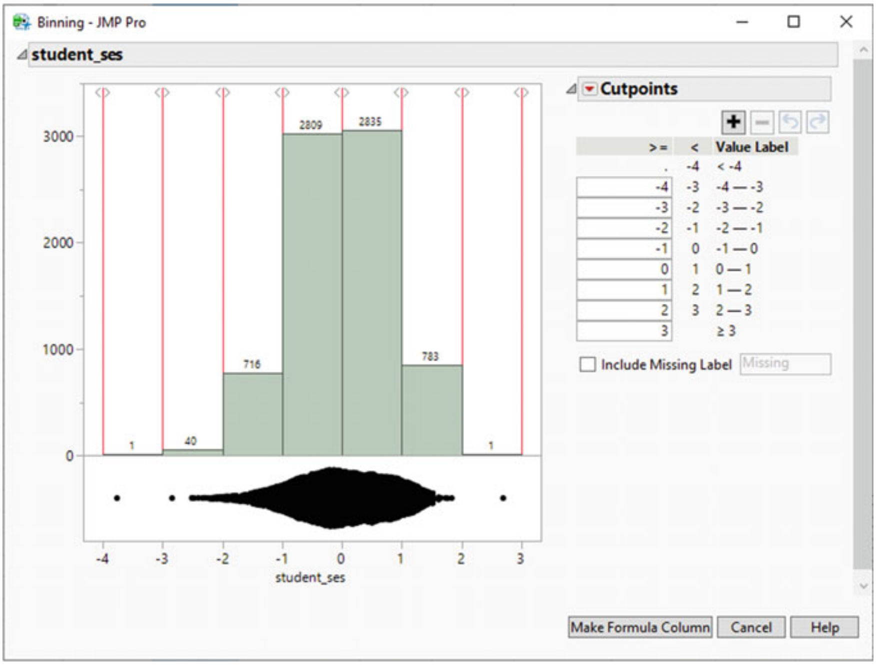


Fig. 5.2 Overplotting in the scatterplot of math achievement and SES

bins, respectively (see Fig. 5.3). Re-assign these two points by entering -4 in the second cell and 3 in the second last cell (see Fig. 5.4).

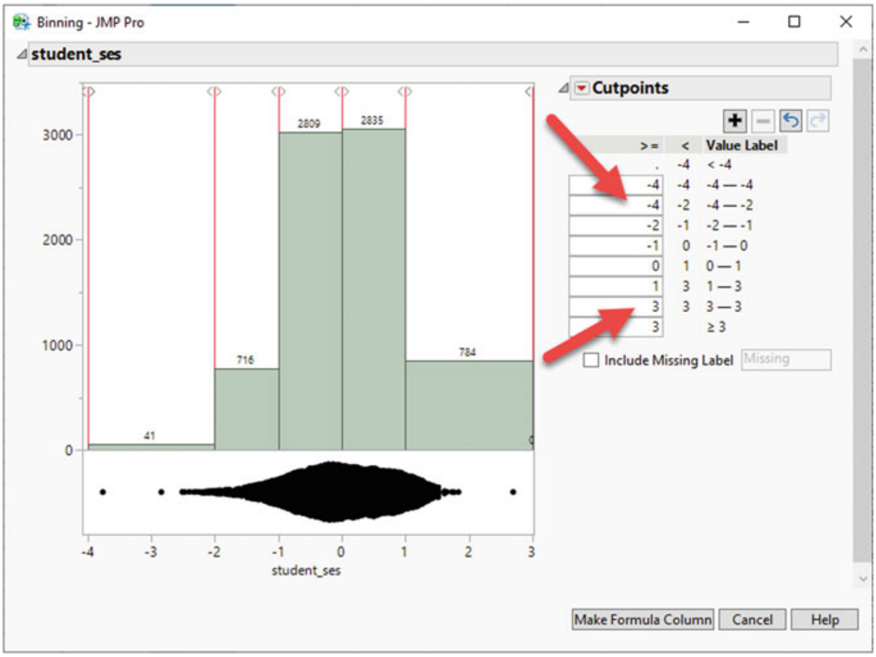
4. Press **Make Formula Column** and a new variable named student_ses_Binned is created.



>=	<	Value Label
.	-4	< -4
-4	-3	-4 — -3
-3	-2	-3 — -2
-2	-1	-2 — -1
-1	0	-1 — 0
0	1	0 — 1
1	2	1 — 2
2	3	2 — 3
3		≥ 3

□ Include Missing Label Missing

Fig. 5.3 Make binning formula for SES



>=	<	Value Label
.	-4	< -4
-4	-4	-4 — -4
-4	-2	-4 — -2
-2	-1	-2 — -1
-1	0	-1 — 0
0	1	0 — 1
1	3	1 — 3
3	3	3 — 3
3		≥ 3

☐ Include Missing Label Missing

Fig. 5.4 Re-assign two observations in making bin formula

Next, the analyst can overcome the problem of overplotting by plotting the median of binned SES against math achievement. The procedures for this data visualization task are as follows (alternatively, the user can run the script "student_mathach vs. student_ses Binned" provided in the data set):

1. Open **Graph Builder** from the pull-down menu **Graph**.
2. Drag student_mathach into the **Y-axis** of the drop zones.
3. Drag student_ses_Binned into the **X-axis**.
4. If the raw data (data points) and a nonlinear fitted line are shown, undo them by clicking on the first two icons from the left (see Fig. 5.5).
5. Press the **boxplot icon** (the ninth one from the left, see Fig. 5.5).

In Fig. 5.5, the median of math achievement is displayed in each bin of SES. Data visualization is essentially data reduction. By hiding the raw data and showing the boxplots only, one can clearly see that as student SES increases, the median of math ability also increases.

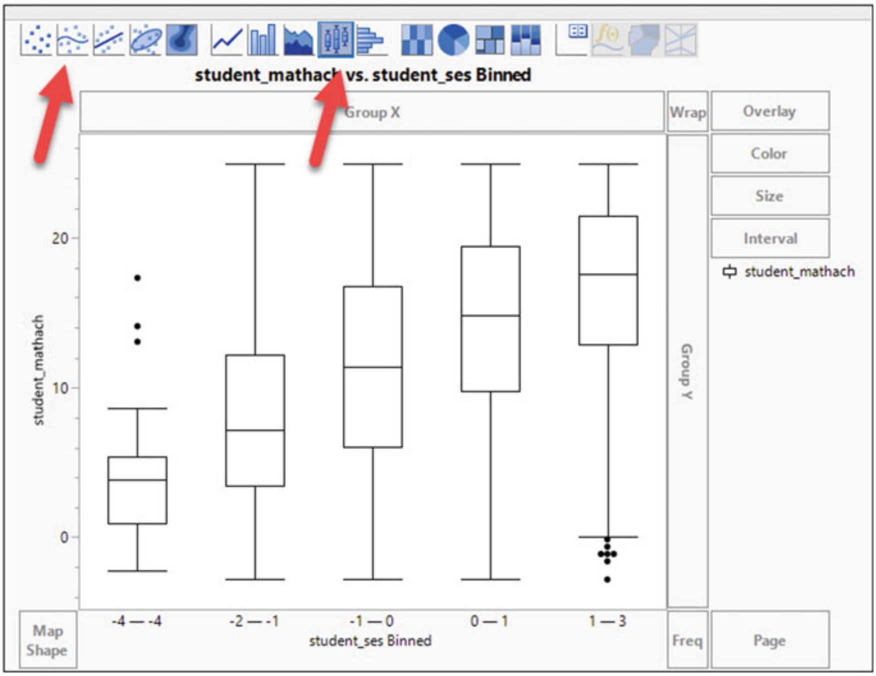


Fig. 5.5 Median smoothing of SES by math achievement

5.3.3 Data Reduction for Examining the Relationship between SES and Math Achievement in Bubble Plot

Besides binning and median smoothing, there are several other data reduction techniques for visualizing complicated relationships, such as the bubble plot (Yu, 2014). As the name implies, in this plot the value of one of the variables (i.e. school size) is depicted by the size of the bubble, and individual students are grouped by their schools. As a result, the graph appears less cluttered, allowing the pattern of the data to easily emerge. To create a bubble plot, the steps below should be followed (the user can also run the script “Bubble Plot of student_mathach by student_ses” in the data set provided):

1. Select **Bubble Plot** from the pull-down menu **Graph**.
2. Drag student_mathach into **Y**.
3. Drag student_ses into **X**.
4. Drag school_ID into **ID**.
5. Drag school_size into **Sizes**.
6. Drag school_sector into **Coloring**.
7. Press **OK**.

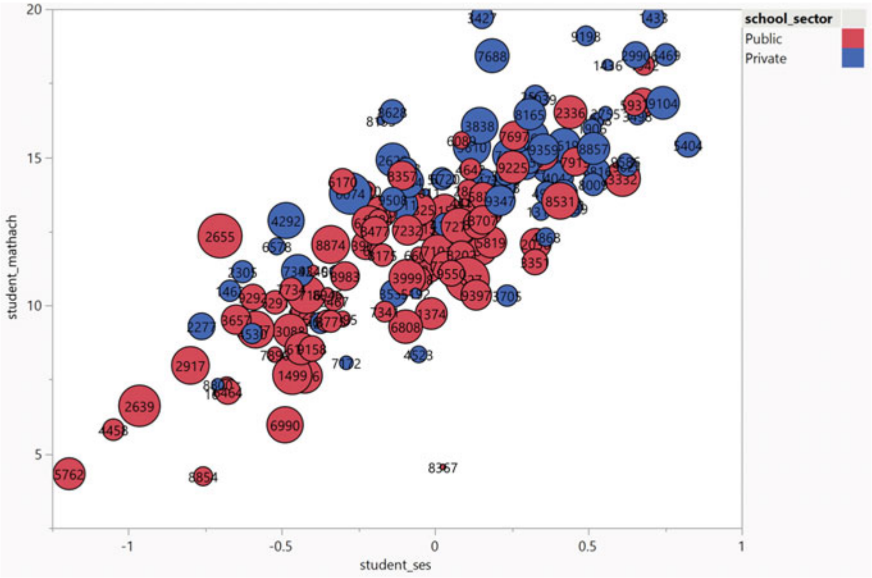


Fig. 5.6 Data reduction by bubble plot

As shown in Fig. 5.6, the problem of over-plotting is alleviated because each data point represents a school. In alignment with the finding indicated using the approach of median smoothing, there is a positive and linear relationship between SES and math ability. Additionally, in concurrence with the finding indicated using the approach of linking and brushing, this relationship does not seem to be affected by school size, because big and small circles scatter around the graph. However, the school sector matters because private schools appear to be concentrated in the upper right-hand corner of the plot. To some extent this is expected, because it is more likely for students from high SES households to attend private schools.

5.3.4 Examining Variation between Schools by ANOM Plot

The preceding exploration sets the stage for further analysis by providing analysts with a global overview of the data. Before going into multilevel modeling, analysts can utilize several data visualization techniques to study school factor, the variable that created the hierarchical structure. The data were sourced from many different schools in the US, which no one would expect to be homogeneous. The question is not whether there are variations between schools in terms of SES and math achievement; rather, the concern is about the degree of variation. Conventional statistical methods, such as multiple comparison procedures, are not suitable for this task because there are too many levels (160 schools). As indicated in Figs. 5.7 and 5.8, straightforward

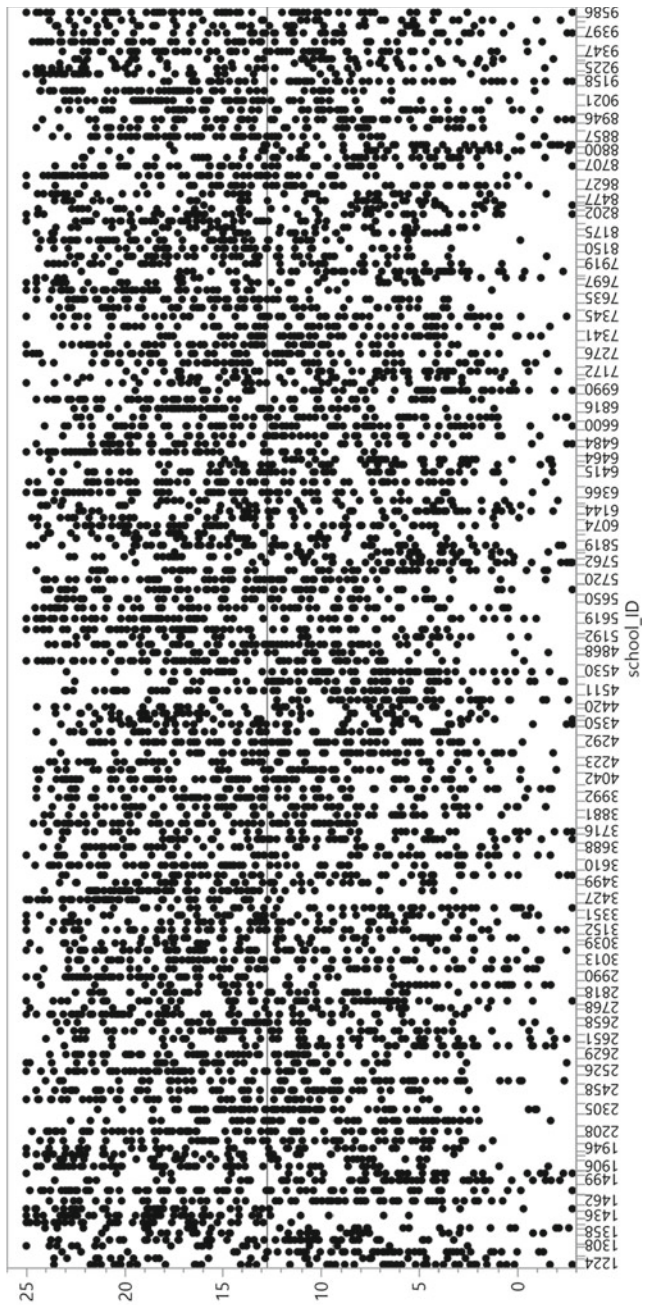


Fig. 5.7 Overplotting of student math achievement by school

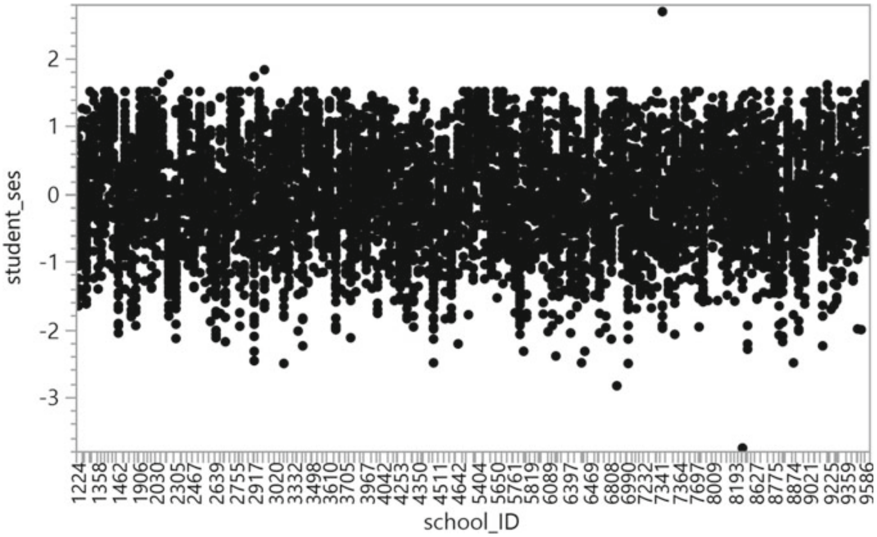


Fig. 5.8 Overplotting of student SES by school

data visualization tools, such as dot plots or histograms, are also problematic because, once again, large sample sizes inevitably leads to overplotting.

In order to visually inspect the degree of variation between schools in terms of math achievement and SES (the two variables of interest), an analyst can utilize a tool named Analysis of Means (ANOM) Plot. ANOM is a graphical and statistical method for simultaneously comparing many means with the grand mean at a specified significance level. The technique was invented for statistical quality control in 1967, and it became popular in manufacturing during the 1980s. Its application spread to the service industry and health care during the 1990s, but it was overlooked by social scientists (Ott, 1967, 1975). Nevertheless, this feature is included in many SAS modules (SAS Institute, 2016). The steps to creating an ANOM chart are (the scripts for this procedure are "ANOM of student_mathach by school_ID" and "ANOM of student_ses by school_ID"):

1. Select **Fit Y by X** from the pull-down menu **Analyze**.
2. Put student_mathach into **Y, Response**.
3. Put school_ID into **X, Factor**.
4. From the inversed red triangle select **Analysis of Means Methods** &#xF0E0; **ANOM**.

Figures 5.9 and 5.10 display which means deviate from the grand mean, thus providing an analyst with information on variations across schools. In Fig. 5.4 the middle gray line represents the grand mean. According to the criteria of the upper and lower decision limits, the red means are considered substantially deviated, whereas the green means are considered acceptable. In other words, any school that is not contained by the upper decision limit (UDL) or lower decision (LDL) is red-flagged.

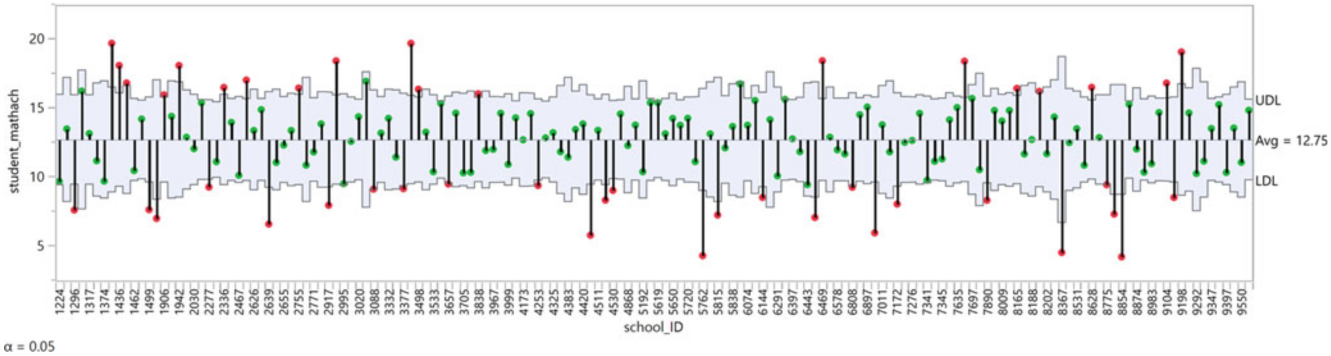


Fig. 5.9 ANOM chart of student math achievement by school

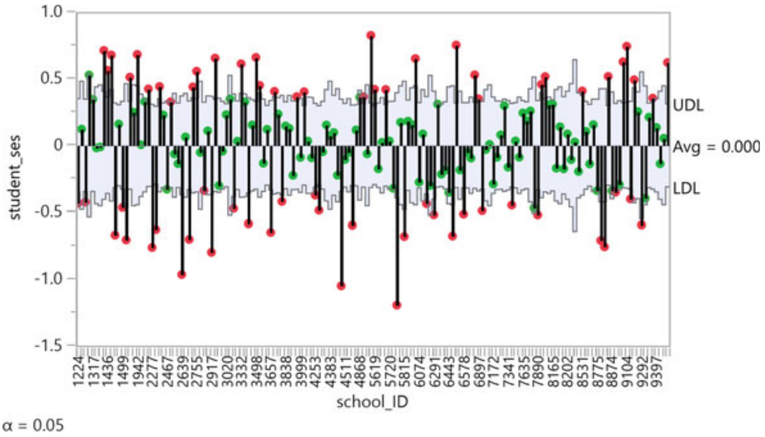


Fig. 5.10 ANOM chart of student SES by school

Each school has its own UDL and LDL, which are determined by: (a) the number of groups, (b) the sample size of each group, (c) the total sample size, (d) the grand mean, (e) the average group variance, and (f) the critical value.

An analyst could go a step further to show a summary report of ANOM. For this purpose only one step is needed: from the inversed red triangle of **Analysis of Means,** select **Show Summary Report**. As shown in Table 5.3, twenty group means exceed the upper limit while twenty-six go beyond the lower one. Overall, 28.13% of 160 schools deviate from the acceptable interval. The variation of SES is even larger (41.88%). Specifically, thirty-four means are above the UDL whereas thirty-three means are below the LDL (To preserve space, the table of SES mean variation is not displayed). Hence, it is reasonable to make a conjecture that the hierarchical structure of the data influences the relationship between SES and math achievement.

5.4 Multilevel Modeling

5.4.1 Computing ICC to Decompose Variance Components by Random Effects Modeling

Conducting data visualization is analogous to computing descriptive statistics before running inferential statistics. Before running a full mixed model, it is beneficial to investigate how much variance in math achievement can be attributed to students and schools by another preliminary analysis, namely, random coefficient modeling (Wolfinger, 1996). The steps for running this model are (the script name is “Random coefficient model using school ID”):

Table 5.3 Group means of math achievement exceeds the UDL or LDL

School ID	*N*	Lower limit	Group mean	Upper limit	Limit exceeded
1433	35	8.95	19.72	16.55	Upper
1436	44	9.36	18.11	16.13	Upper
1461	33	8.84	16.84	16.66	Upper
1906	53	9.67	15.98	15.83	Upper
1942	29	8.57	18.11	16.92	Upper
2336	47	9.47	16.52	16.02	Upper
2526	57	9.78	17.05	15.72	Upper
2755	47	9.47	16.48	16.02	Upper
2990	48	9.51	18.45	15.99	Upper
3427	49	9.54	19.72	15.95	Upper
3498	53	9.67	16.39	15.83	Upper
3838	54	9.69	16.06	15.80	Upper
6469	57	9.78	18.46	15.72	Upper
7688	54	9.69	18.42	15.80	Upper
8165	49	9.54	16.45	15.95	Upper
8193	43	9.32	16.23	16.17	Upper
8628	61	9.88	16.53	15.62	Upper
9104	55	9.72	16.83	15.77	Upper
9198	31	8.71	19.09	16.78	Upper
1296	48	9.51	7.64	15.99	Lower
1499	53	9.67	7.66	15.83	Lower
1637	27	8.42	7.02	17.07	Lower
2277	61	9.88	9.30	15.62	Lower
2639	42	9.28	6.62	16.21	Lower
2917	43	9.32	7.98	16.17	Lower
3088	39	9.15	9.15	16.34	Lower
3377	45	9.40	9.19	16.09	Lower
3657	51	9.61	9.52	15.89	Lower
4253	58	9.80	9.41	15.69	Lower
4458	48	9.51	5.81	15.99	Lower
4523	47	9.47	8.35	16.02	Lower
4530	63	9.92	9.06	15.57	Lower
5762	37	9.05	4.32	16.44	Lower
5815	25	8.25	7.27	17.24	Lower
6144	43	9.32	8.55	16.17	Lower
6464	29	8.57	7.09	16.92	Lower

(continued)

Table 5.3 (continued)

School ID	*N*	Lower limit	Group mean	Upper limit	Limit exceeded
6808	44	9.36	9.29	16.13	Lower
6990	53	9.67	5.98	15.83	Lower
7172	44	9.36	8.07	16.13	Lower
7890	51	9.61	8.34	15.89	Lower
8367	14	6.73	4.55	18.76	Lower
8775	48	9.51	9.47	15.99	Lower
8800	32	8.78	7.34	16.72	Lower
8854	32	8.78	4.24	16.72	Lower
9158	53	9.67	8.55	15.83	Lower

1. Choose **Fit Models** from **Analyze**.
2. Choose **Mixed Model** from **Personality** at the upper right corner of the pop-up window.
3. Put student_mathach into **Y**.
4. Click on the tab **Random Effects** on **Construct Model Effect**.
5. Add school_ID into **Random Effects**.
6. Click **Run**.

In the output there is a table entitled "Random Effects Covariance Parameter Estimates," as shown in Table 5.4. According to Table 5.4, the estimate of the variance between schools is 8.61 while that of the residual is 47.76. The *p*-value of the school effect is < 0.0001, suggesting that the factor of school does make a difference on math achievement. An analyst can employ the intraclass correlation coefficients (ICC) to decompose the variance components. An ICC is commonly used for estimating inter-rater reliability and test–retest reliability. Specifically, when different raters grade the same tests, a psychometrician wants to know whether different raters agree with one another. By the same token, when the same students retake the same test repeatedly, a psychometric researcher wants to affirm that the results are stable over time. Put otherwise, this type of agreeableness or stability is a form of relatedness or dependence. In multilevel modeling, an ICC can be employed as a measure of dependence between subjects. The formula of ICC is: σ^2effect/(σ^2effect + σ^2residual), and thus 8.61/(8.61 + 39.15) = 0.18. Based on the random effects model, it can be concluded that 18% of

Table 5.4 Random effects covariance parameter estimates for ICC

Variance component	Estimate	Std error	95% lower	95% upper	Wald *p*-value
school_ID	8.61	1.08	6.49	10.73	< 0.0001*
Residual	39.15	0.66	37.89	40.48	
Total	47.76	1.26	45.39	50.32	

* Significant at the 0.01 level

the variance in math achievement scores is due to between-school differences while 82% can be explained by within-school differences.

5.5 Running a Mixed Model for Fixed and Random Effects using SES

Let's revisit the difference between fixed and random factors in a multilevel model. When a variable contains all the levels that a researcher cares about, it is considered a fixed factor. In this example, SES ranges from the upper class to the lower class, and no other possible value is out of that range. In contrast, if a variable only carries a subset of all possible levels, then it is considered a random factor. In this example, even though HSB is a national study, the data set does not include every school in the US. Hence, it is treated as a random sample of all American high schools. The steps to building a mixed model for nested data are (see Fig. 5.11; the script is "Mixed model using SES centered by grand mean"):

1. Choose **Fit Model** from **Analyze**.
2. At the upper right corner of the pop-up window, select **Mixed Model** from **Personality**.
3. Add student_mathach into **Y**.
4. Add student_ses into **Fixed Effects**.
5. Click on the tab **Random Effects**.
6. Add student_ses into **Random Effects** and then click on student_ses.
7. Click on school_ID in **Select Columns**.

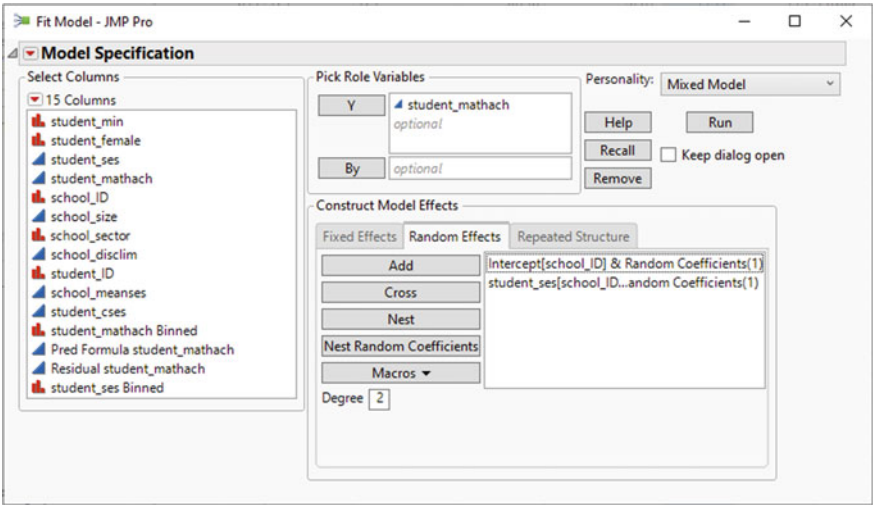

Personality: Mixed Model
Y: student_mathach
Random Effects:
Intercept[school_ID] & Random Coefficients(1)
student_ses[school_ID...andom Coefficients(1)
Degree 2

Fig. 5.11 Running mixed model in JMP

8. In **Construct Model Effects** press **Nest Random Coefficients**.
9. Press **Run**.

Figure 5.12 is a plot of actual math score against the predicted score, in which the model fit is accounted by the fixed effect. Figure 5.13 is similar to Fig. 5.12 except that it is a conditional plot, meaning that the relationship between SES and math competency is conditional by school. This plot shows a visual assessment of model fit that accounts for variation attributed to random effects. These fits are reasonable because only SES is taken into account, while all other factors (e.g. minority status, gender, school sector, etc.) are set aside. As indicated by linking/brushing and usage of the aforementioned bubble plot, an analyst can consider including other potential predictors in the future inquiry for improving the model fit.

Model fitness can be further examined by checking various residual plots, which are generated by choosing residual plots from **Marginal Model Inferences** under

Fig. 5.12 Plot of actual and predicted math score

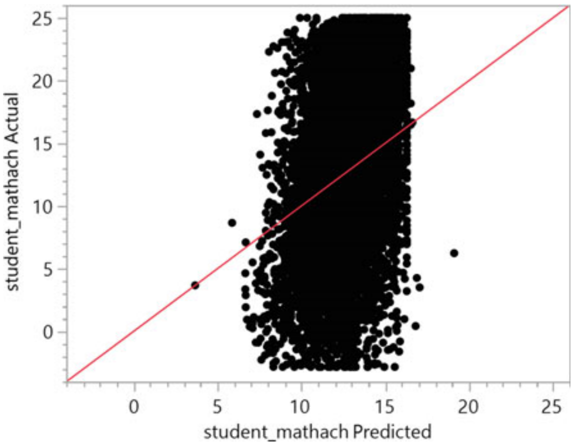
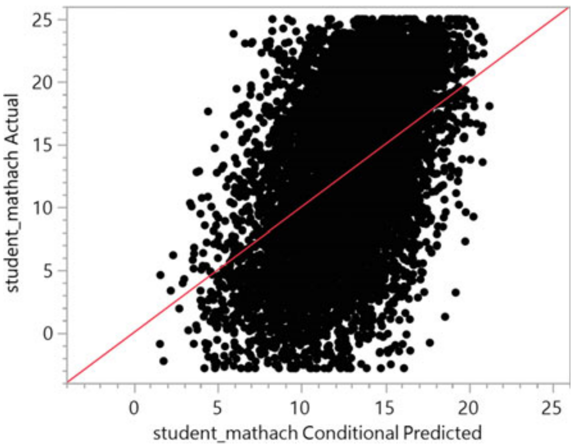


Fig. 5.13 Conditional plot of actual and predicted math score



the inversed red triangle. In Fig. 5.14, the histogram located at the lower right-hand corner indicates that the residuals are normally distributed whereas the residual by row plot at the lower left hand side shows that the residuals randomly scatter around zero. However, it is unsettling to see a "cliff" in the residual by predicted plot at the upper left corner (see the red arrow). It seems that most predicted values of student math achievement hit a ceiling and also there are several extreme residuals at the two polarities of student_mathach predicted. This phenomenon can be attributed to the fact that local information of each school is partially suppressed by the grand mean of SES.

In the fixed effect output (Table 5.5) the intercept is 12.67, meaning that for a typical student with a SES value of 0 (SES is centered by the grand mean at zero), the gain in the math score is 12.67. The estimate of the within-effect of SES is 2.39,

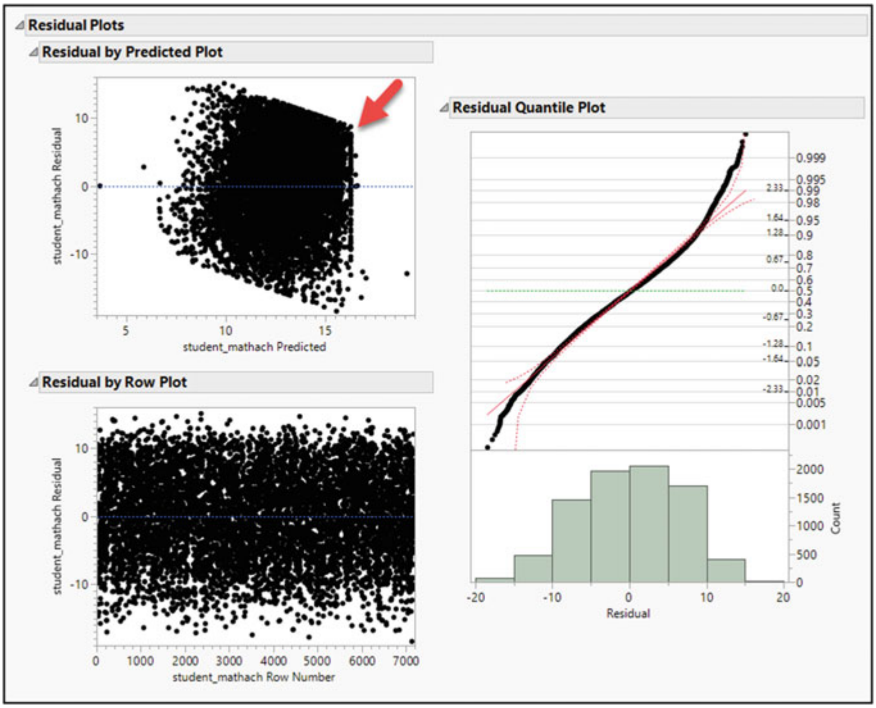


Fig. 5.14 Residual plots of using student SES to predict math achievement in mixed modeling

Table 5.5 Fixed effects parameter estimates

Term	Estimate	Std error	DF	t ratio	Prob > |t|	95% lower	95% upper
Intercept	12.67	0.19	145.6	66.66	< 0.0001*	12.29	13.05
student_ses	2.39	0.12	157.5	20.22	< 0.0001*	2.169	2.63

* Significant at the 0.01 level

Table 5.6 Random effects covariance parameter estimates

Covariance parameter	Subject	Estimate	Std error	95% lower	95% upper	Wald *p*-value
Var(Intercept)	school_ID	4.83	0.67	3.5	6.15	< 0.0001*
Cov(Intercept,student_ses)	school_ID	-0.15	0.29	-0.74	0.43	0.6057
Var(student_ses)	school_ID	0.41	0.24	-0.05	0.87	0.0789
Residual		36.83	0.63	35.63	38.09	

* Significant at the 0.01 level

but it is important to point out that this fixed effect results from lumping together observations from all schools.

Table 5.6 shows the random effects covariance parameter estimates, which includes the intercept and slope variances, as well as the covariance with Wald confidence intervals. The confidence intervals indicate the variance of SES; the covariance is not substantially different from zero, while the intercept (4.83 with a SE of 0.67) is significant. It is expected because the preliminary analysis using an ANOM chart shows a wide variability of math achievement across different schools.

5.5.1 Using School Mean and Locally Centered SES to Disentangle Within and between Groups

As mentioned in the section "Data Source," the original student SES is centered by the grand mean. One may argue that this type of centering ignores the local context of each school; to be more precise, the SES of the students should be centered by the SES mean of the school they attend. Consider this analogy: if the household income data obtained in Montana are centered by the national average, the results could be misguiding because it doesn't take the local living standard into account. To adjust the mean by local information, another variable named student_cses is included in the data set. Further, after all, that is the school level creating the hierarchical structure, and thus it is appropriate to include the SES mean of each school (school_meanses) as a predictor. One might wonder why the original SES centered by the grand mean is retained if the new SES centered by the local mean seems to be more precise. There are pros and cons with the latter approach. Although the use of SES centered by the school mean is more precise, this variable cannot provide a researcher with information on between-school dispersion. Figure 5.15 is an ANOM plot of student_cses by school (The user can run the script "ANOM of student_cses by school_ID" to generate the graph). After adjusting the student SES by the local mean, no school shows any deviation. Needless to say, this is misleading. Thus, it is advisable to start multilevel analysis with the SES centered by the grand mean, and subsequently move into more sophisticated modeling.

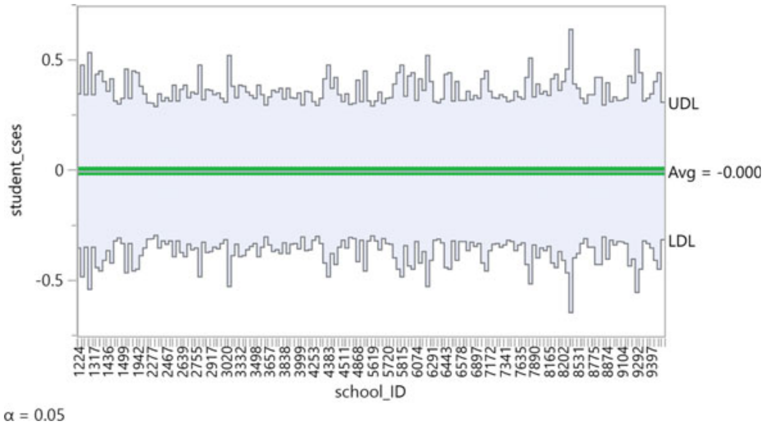


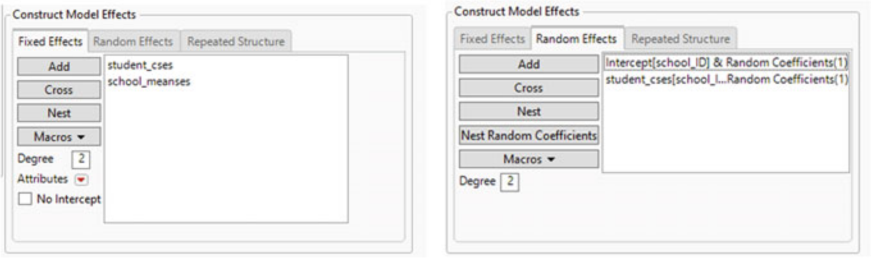
Fig. 5.15 ANOM plot of student_cses by school

To run a fine-tuned mixed model, both student_cses and school_meanses should be included in Fixed Effects. For the random effect, student_cses is nested with school_ID (see Fig. 5.16; the script is "Mixed model using cses and school mean").

When the residual by predicted model is examined, a different story emerges. Unlike Fig. 5.14, this residual plot in Fig. 5.17 no longer shows a "cliff" or extreme residuals, meaning that the model has been improved by taking the variance between schools and local information of each school into account.

The estimate of the within-school effect of SES was 2.39 (see Table 5.7) when student_ses was used as the fixed factor. In this modified model, the estimate is shrunken to 2.19 and the estimate of the between-school effect of SES is 5.89. By doing so the within-component and between-component of variation in SES is disentangled.

No matter whether student_ses or student_cses is used as a predictor, JMP produces sub-models for different schools. Table 5.8 is a partial list of the intercept and the regression coefficient of each school when student_cses is used for modeling. There are 160 schools in this data set. To save space, only the first 10


Construct Model Effects

Fixed Effects | Random Effects | Repeated Structure

student_cses
school_meanses

Degree 2
□ No Intercept

Construct Model Effects

Fixed Effects | Random Effects | Repeated Structure

Intercept[school_ID] & Random Coefficients(1)
student_cses[school_I...Random Coefficients(1)

Degree 2

Fig. 5.16 Construct a multilevel model with school ses mean and student SES centered by local mean

Fig. 5.17 Residual by predicted plot using school mean and student SES centered by each local mean to predict math achievement in mixed modeling

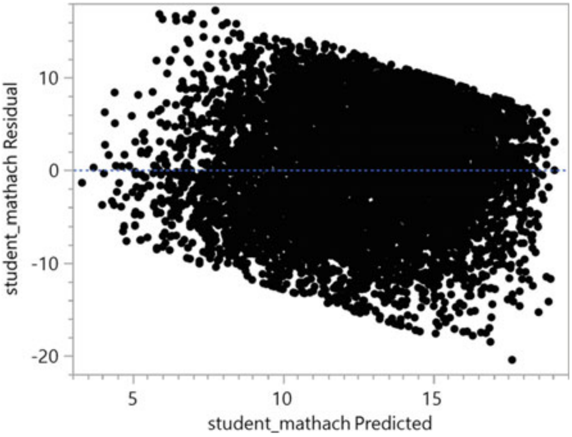


Table 5.7 Fixed effects parameter estimated using locally centered SES and school SES mean

Term	Estimate	Std error	DF	t ratio	Prob > |t|	95% lower	95% upper
Intercept	12.69	0.15	153.7	84.96	< 0.0001*	12.39	12.98
student_cses	2.19	0.13	155.4	17.10	< 0.0001*	1.939	2.448
school_meanses	5.89	0.36	152.9	16.28	< 0.0001*	5.189	6.618

* Significant at the 0.01 level

Table 5.8 Intercept and random coefficients for different schools

school_ID	Intercept	student_cses
1224	−0.32	0.10
1288	0.05	0.17
1296	−1.94	−0.157
1308	0.29	−0.18
1317	−1.18	−0.09
1358	−0.99	0.61
1374	−1.97	0.48
1433	2.05	-0.23
1436	1.62	−0.25
1461	0.02	0.88

are shown. The full table provides 160 sub-models, as opposed to a single "one size fits all" model. This random coefficient table reports the best linear unbiased predictor (BLUP) values (Henderson, 1975) for how each school differs from the population intercept and population SES effect in **Fixed Effects Parameter Estimates**. The BLUP parameter is an estimate of the random effect least squares mean, which has a tendency of shrinking toward the grand mean, thus minimizing mean square prediction error. The amount of shrinkage is tied to the variance of the effect and the number of observations of the level. When the variance component is large, there is little to no shrinkage. Conversely, if the variance is small, the shrinkage is large. If the variance component is zero, the result of shrinkage is exactly zero. In this sense, a fixed effect can be considered a special case of the random effect when the variance component is very large and it leans toward the grand mean (Cao, 2015, Hummel et al., 2021, SAS Institute, 2020). Put it another way, BLUP can also be conceptualized as a form of empirical Bayes estimator (Robinson, 1991). In a classical Bayesian analysis, usually the analyst subjectively set a prior estimate and then the posterior probability is updated by data. As the name implies, the empirical Bayes estimator is data-driven. In contrast to the classical Bayesian approach, in this case the prior is estimated from all the data (e.g. all schools) and subsequently an update of coefficients is made to each school based on the local evidence.

5.5.1.1 Using Local Data Filter to Examine Submodels

It is important to note that the coefficients of BLUP cannot reveal the data pattern or answer any of the following questions: Are there outliers in a particular school? When all 7,185 observations are utilized for data analysis, a few outliers would not influence the outcome. There are between 14 and 67 students per school. In a much smaller subset outliers can become influential points. Further, is there a curvilinear relationship in a particular school? Is there any clustering pattern in a particular school? To answer the preceding question, it is imperative to create a conditional scatterplot, a graph that displays the relationship between SES and math achievement conditioned on the factor of school. The steps for creating a conditional plot is as follows (the script is “Conditional scatterplot of student_mathach vs. student_ses all”):

1. Open Graph Builder from the pull-down menu Graph.
2. Drag student_mathach into the **Y-axis**.
3. Drag student_ses into the **X-axis**.
4. Drag school_ID into **Overlay**.

It seems that overplotting would hinder an analyst’s ability to extract any meaningful pattern from the data (see Fig. 5.18). While most regression lines are positive, some of them appear to be negative or flat. Once again, dynamic graphics allow the analyst to explore the data in an interactive fashion.

When the analyst clicks on one of the negative lines and one of the flat ones, only these two lines are active and all others are grayed out. In addition, the labels of the school IDs associated with these two lines are italicized: School 2277 and 2629 (see Fig. 5.19).

Utilizing a local data filter, the analyst can examine the bivariate relationship for different schools. The procedures for adding a local data filter and selecting a particular data subset are (the script is “Local data filter for student_mathach vs. student_ses”):

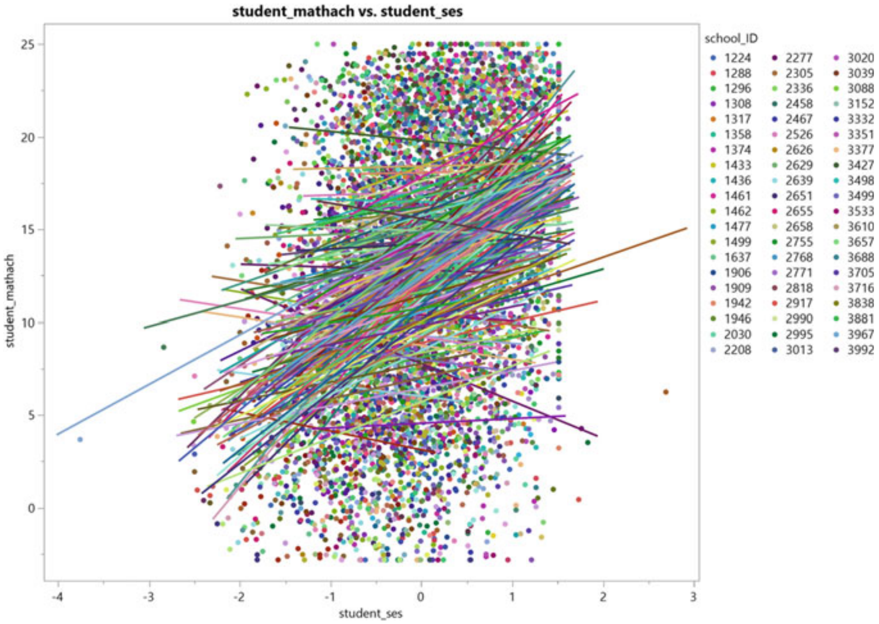


Fig. 5.18 Scatterplot showing the relationship between SES and math achievement conditioning on school

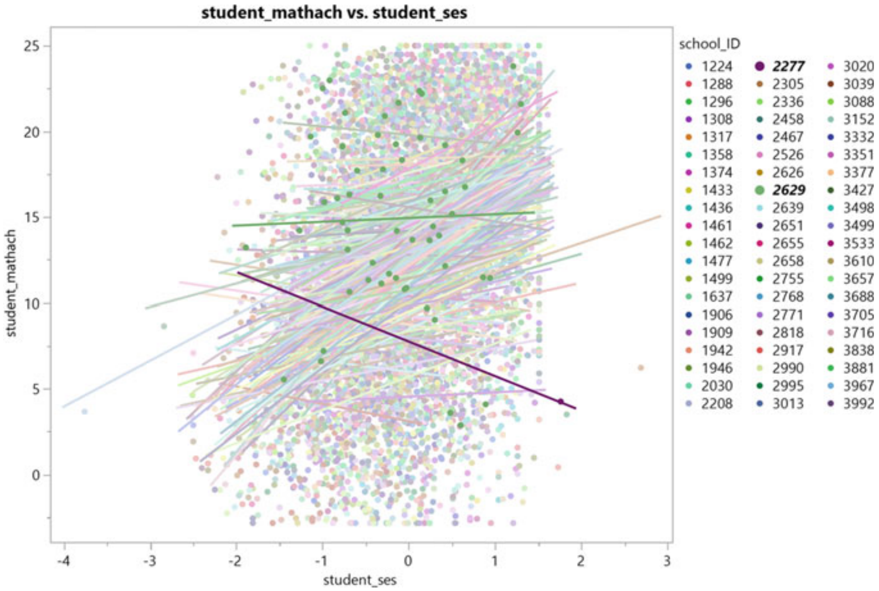


Fig. 5.19 Two schools are selected in conditional scatterplot

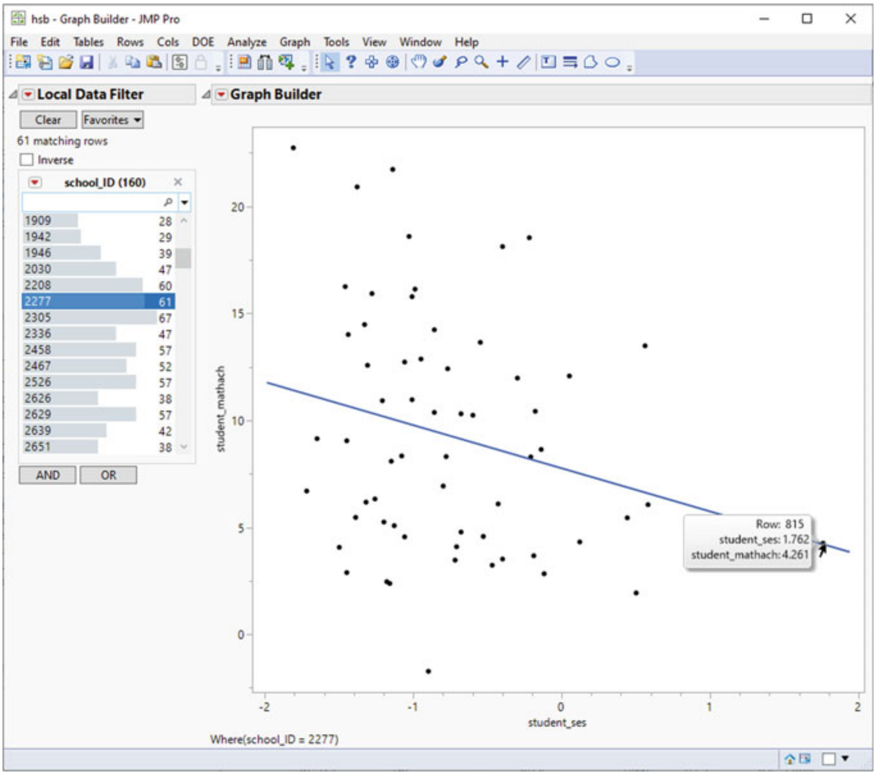
1. From the first inversed **red triangle** next to Graph Builder select **Local Data Filter**.
2. From **Local Data Filter** select school_ID.
3. Scroll down and select 2277.

As shown in Fig. 5.20, all other data are filtered out and the submodel contains only 61 students from School 2277. There appears to be an outlier located at the lower right-hand corner of the graph. When the cursor hovers upon a data point, a pop-up window displays its row number and values of SES and math competency.

An analyst can follow these steps to hide and exclude this outlier (see Fig. 5.21):

1. Right-click on Row 815 in the table.
2. Select **Hide and Exclude** and the graph can be automatically updated.

The updated graph (see Fig. 5.22) shows that the regression slopes have changed slightly, but the relationship is still negative. The analyst can click on School 2629 and a different school ID in the local data filter to continue to explore different sub-models by school.


hsb - Graph Builder - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Local Data Filter
Graph Builder
Clear
Favorites
AND
OR
20
15
10
5
0
-2
-1
0
1
2
student_mathach
student_ses

61 matching rows

☐ Inverse

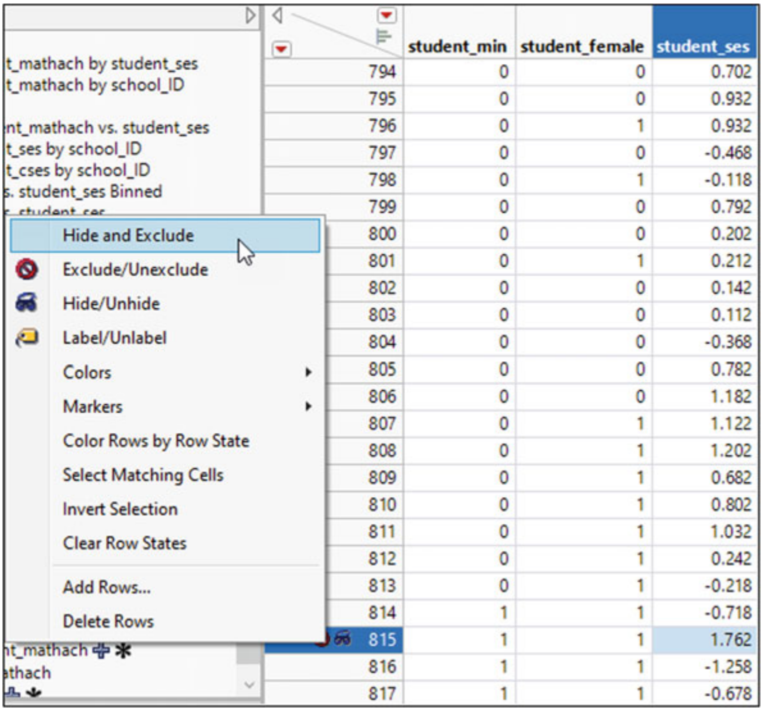
school_ID (160)

1909	28
1942	29
1946	39
2030	47
2208	60
2277	61
2305	67
2336	47
2458	57
2467	52
2526	57
2626	38
2629	57
2639	42
2651	38

Row: 815
student_ses: 1.762
student_mathach: 4.261

Where(school_ID = 2277)

Fig. 5.20 Using local data filter to examine the submodel of school 2277



	student_min	student_female	student_ses
794	0	0	0.702
795	0	0	0.932
796	0	1	0.932
797	0	0	-0.468
798	0	1	-0.118
799	0	0	0.792
800	0	0	0.202
801	0	1	0.212
802	0	0	0.142
803	0	0	0.112
804	0	0	-0.368
805	0	0	0.782
806	0	0	1.182
807	0	1	1.122
808	0	1	1.202
809	0	1	0.682
810	0	1	0.802
811	0	1	1.032
812	0	1	0.242
813	0	1	-0.218
814	1	1	-0.718
815	1	1	1.762
816	1	1	-1.258
817	1	1	-0.678

Fig. 5.21 Hide and exclude an outlier

5.6 Conclusion

Today many analytical tools are available for multilevel modeling. However, educational researchers often overlook pattern seeking. The key difference between regular data sets and nested data sets is in their data structure. As such, it is imperative to examine data patterns in order to obtain a holistic perspective on subject matters under study. Linking and brushing, which is a vital feature of dynamic visualization, provides researchers with a global overview of interrelationships between all available variables. When an analyst would like to investigate relationships between focal dependent and independent variables, the problem of overplotting arises in large samples. There exist different approaches to overcoming this obstacle. Binning and median smoothing reduce data into a few intervals so that the data trend can be detected by checking the median of math achievement at a few levels of SES. As a multivariate data visualization technique, the bubble plot summarizes data by school and also displays interrelationships among several variables. An ANOM plot enables an analyst to visually examine variations between schools in student math ability and SES. All of the above preliminary analyses can be used to gather background information and set the stage for multilevel modeling.

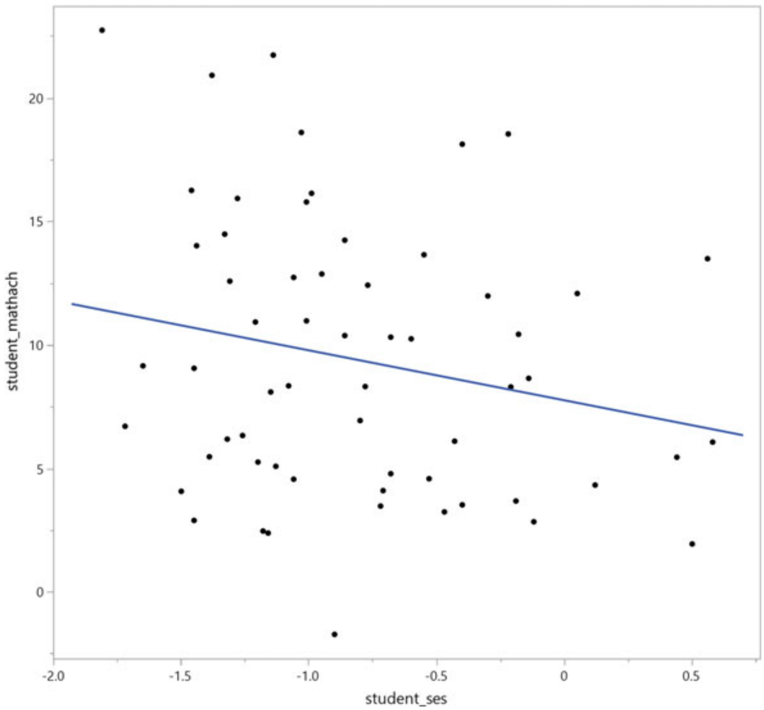


Fig. 5.22 Updated scatterplot after hiding and excluding an outlier

To decompose variances of students and schools, an ICC can be obtained from a simple random effect model. In mutlilevel modeling an analyst can construct a model using the SES centered by the grand mean or one centered by the local school mean. Either model yields BLUP, the random coefficients for different schools. However, these numbers alone cannot inform an analyst about the pattern in each submodel. To rectify this situation, an analyst can partition the results and examine each school, using the local data filter. The preceding methods are by no means exhaustive and readers are encouraged to employ other creative ways to scrutinize hierarchical data.

More importantly, for ease of illustration the data set used in this chapter has two levels only (student and school). As mentioned in the beginning, it is common for an international or national study to include multiple levels, such as state, county, school district, school, and student. Nonetheless, the same logic of decomposing variance components can be well applied to more complicated nested data. Further, when the sample size is extremely large (e.g. the data set is both multilevel and longitudinal), running a mixed model on desktop would tie up all the system's resources. To rectify the situation, it is advisable to utilize in-memory analytics and multi-threaded processing by moving from desktop computing to cloud computing, such as SAS Viya (SAS Institute, 2020). Specifically, rather than downloading a huge data set from the data warehouse, the cloud-based system can perform analysis while the data are still held in the computer memory. When multi-threaded processing is used, the processor can handle high performance computing by partitioning and analyzing the data in multiple threads concurrently. Researchers who utilize big archival data are encouraged to explore this option (Gottula, 2021).

References

Bauer, D., & Curran, P. (2017). *Multilevel modeling of hierarchical and longitudinal data using SAS*. SAS Institute.

Bosky, A.L. (2019). *Academic preparation in high school and gendered exposure to economic insecurity at midlife* (Publication no. 28219447) [Doctoral dissertation, University of Texas at Austin]. ProQuest Dissertations and Theses Global.

Cao, J. (2015, March). *Linear mixed model with JMP Pro: One face, many flavours*. https://community.jmp.com/kvoqx44227/attachments/kvoqx44227/discovery-eu-2015-content/28/2/Linear%20Mixed%20Models%20With%20JMP%20Pro%20One%20Face.pdf

Carr, D. B. (1991). Looking at large data sets using binned data plots. In A. Buja & P. A. Tukey (Eds.), *Computing and graphics in statistics* (pp. 5–39). Springer-Verlag.

Gottula, J. (2021, July 12). *SAS tutorials: SAS for mixed models* [Video]. YouTube. https://www.youtube.com/watch?v=by9hhoDpBIk&list=PLVV6eZFA22QwrXd6nSDU18E6XgXSMOs87

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics, 31*(2), 423–447.

Hummel, R. M., Claassen, E. A., & Wolfinger, R. D. (2021). *JMP for mixed models*. SAS Institute.

National Center for Educational Statistics [NCES]. (2018). *Measuring school climate using the 2015 school crime supplement: Technical report*. Author. https://nces.ed.gov/pubs2018.pdf

National Center for Educational Statistics [NCES]. (2020). *NCES Handbook of survey method*. https://nces.ed.gov/statprog/handbook/hsb_keyconcepts.asp

National Center for Educational Statistics [NCES]. (2021). *High school and beyond*. https://nces.ed.gov/surveys/hsb/

Ott, E. R. (1967). Analysis of means: A graphical procedure. *Industrial Quality Control, 24*, 101–109.

Ott, E. R. (1975). *Process quality control: Troubleshooting and interpretation of data*. McGraw-Hill.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with rejoinder). *Statistical Science, 6*(1), 15–32.

SAS Institute. (2016). *SAS/QC 14.2 user's guide*: *The ANOM procedure*. Author.

SAS Institute. (2020). *SAS visual analytics for SAS Viya*. Author.

SAS Institute. (2021). *JMP Pro* (Version 16) [Computer software]. http://www.jmp.com.

Thompson, S. K. (2012). *Sampling*. Wiley.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company.

Wolfinger, R. D. (1996). Heterogeneous variance covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1*(2), 205–230.

Yu, C. H. (2014). *Dancing with the data: The art and science of data visualization*. Lambert Academic Publishing.

Yu, C. H. (2017). Exploratory data analysis. In D. Bricken (Ed.). *Oxford bibliographies*. Oxford University Press. https://doi.org/10.1093/OBO/9780199828340-0200, https://www.oxfordbibliographies.com/view/document/obo-9780199828340/obo-9780199828340-0200.xml?rskey=JOlkN9&result=50

Yu, C. H., & Behrens, J. T. (1995). Applications of scientific multivariate visualization to behavioral sciences. *Behavior Research Methods, Instruments, and Computers, 27*, 264–271.

Chong Ho Alex Yu holds a Ph.D. in Educational Psychology (Arizona State University, ASU) focusing on Measurement, Statistics, and Methodological Studies and a Ph.D. in Philosophy (ASU) with a specialization in Philosophy of Science. Currently, he is a Professor of Behavioral and Applied Science and a member of the adjunct mathematics faculty at Azusa Pacific University (APU), United States. He is also a quantitative research consultant and the Big Data Discovery Summit/Data Science Consortium committee chair at APU. His research interests include alternate research methods (e.g., data visualization, data mining) and cross-cultural comparison (e.g., PISA, TIMSS).

Chapter 6
Doubly Latent Multilevel Structural Equation Modeling: An Overview of Main Concepts and Empirical Illustration

Irena Burić

Abstract Many previous studies that investigated the association between classroom climate variables and student outcomes have suffered from methodological flaws, such as relying exclusively on self-reports, failing to consider the appropriate level of statistical analysis, and inadequately controlling for potential measurement and sampling errors. Such analytical strategies typically lead to confounding the true effects at student and classroom levels as well as to biased estimates. The present chapter provides an overview of main concepts of doubly latent multilevel structural equation modeling (DL-MSEM) that enables testing theoretically relevant relationships at proper level of analysis (i.e., class, teacher, school) and controlling for measurement (by using multiple indicators for latent variables at student and teacher levels) and sampling errors (by incorporating the scores for different students in the same class as multiple indicators of latent variables at teacher level). In addition, an empirical illustration of data analysis with the DL-MSEM is provided by using data based on multilevel design (i.e., students nested within teachers) and drawn from multiple sources (i.e., teachers' and students' reports). More specifically, to assess the climate effects, each student within a class directly rated the instructional behavior of their teacher, thus making a teacher (rather than a student) the referent. In addition, teacher self-reports of their personal characteristic were combined with student reports of their teachers' instructional behavior and student self-reports of their motivational processes. The results were interpreted in relation to the main concepts of the DL-MLSEM method.

Keywords Doubly latent multilevel structural equation modeling · Measurement and sampling errors · Multilevel mediation · Climate vs. contextual effects · Teachers

This research was supported by Croatian Science Foundation (Grant No. UIP-11–2013-5065).

I. Burić (✉)
University of Zadar, Zadar, Croatia
e-mail: iburic@unizd.hr

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_6

6.1 Introduction

An inherent characteristic of most data in educational psychology research is its hierarchical structure—students are nested in classrooms, classrooms are nested in schools, schools are nested in neighborhoods, etc. Such hierarchical data structure implies possible existence of variance at each of these levels. Failure to properly model the variance at the level it naturally exists (i.e., by applying ordinary single-level regression analysis), violates the assumption of independence of observations and often leads to downwardly biased standard errors and, consequently, to erroneous conclusions. Traditional multilevel modeling (MLM) approach (also called hierarchical linear modeling (HLM), mixed modeling, or random-coefficient modeling), to analyzing hierarchically structured (or nested) data surpasses this issue by partitioning the variance of the dependent variable in its within-cluster (level-1) and between-cluster (level-2) parts and flexibly handling predictor variables at any level of measurement (Snijders & Bosker, 2011). In addition, MLM allows the analysis of multilevel data with more accurate Type I error control when compared to the traditional regression methods (Raudenbush & Bryk, 2002).

Nonetheless, MLM is limited in several ways—multivariate models are in most cases difficult to specify, dependent variable can exist only at the lowest level of analysis, most model fit indices are not available, and latent variables are nearly impossible to include in the model, thus, resulting in attenuation of effect sizes due to measurement error. The latter issue, along with sampling error, is especially important when it comes to level-2 constructs aggregated from the level-1 individual ratings since it can introduce bias while estimating the level-2 effects (Jia & Konold, 2021). In addition, in MLM, effects at different levels of analysis are conflated since effects of level-1 predictors are not properly decomposed into within-cluster and between-cluster effects causing level-1 slopes to be weighted average of the between- and within-cluster effects of the level-1 predictor (Preacher et al., 2010). Moreover, steps taken to decompose the effects of the level-1 predictor can also lead to bias (Lüdtke et al., 2008; Preacher et al., 2010). In contrast to MLM, structural equation modeling (SEM) easily incorporates complex models and multivariate outcomes as well as controls for measurement error by including latent variables. However, like traditional regression methods, it cannot properly handle clustered or hierarchically organized data (due to violation of independence of observation assumptions), thus, leading to untrustworthy results in terms of parameter estimates, standard errors, and model fit.

Multilevel structural equation modeling (MSEM) framework (Muthèn & Asparouhov, 2011; Preacher et al., 2010) seems to overcome these limitations by combining the best features of SEM and MLM. Specifically, MSEM allows an investigation of complex relationships between variables while controlling for measurement error, which is central to SEM, with an emphasis on micro–macro relationships, which is central to MLM (Bovaird & Shaw, 2012; Silva et al., 2019). MSEM decomposes the observed level-1 ratings into two orthogonal parts (i.e., level-1 and level-2 variance components) thereby separating the level-1 and level-2 effects and removing bias when estimating level-2 effects. As such, MSEM proved to be highly flexible analytical tool that enables testing theoretically diverse and complex models at different levels of analysis while simultaneously obtaining true (unbiased) effects, controlling for measurement error, and including outcomes that naturally exist only at higher levels (Muthèn & Asparouhov, 2011). Moreover, MSEM demonstrated to be especially useful for examination of contextual and climate effects in educational psychology research (Jia & Konold, 2021; Marsh et al., 2009; Morin et al., 2014) as well as for testing mediating and moderating effects based on various types of multilevel designs (Preacher et al., 2010, 2016; Zyphur et al., 2019).

Thus, in the present chapter, I will demonstrate the application of so-called doubly latent multilevel structural equation modeling (DL-MLSEM; Lüdtke et al., 2011; Marsh et al., 2012; Morin et al., 2021), which can be viewed as a special form of MSEM, in analyzing relationships between teacher self-efficacy (TSE), instructional quality, and student motivational outcomes. The data were drawn from the study of Burić and Kim (2020) and included reports from 94 high school teachers who gave reports on their TSE and their 2087 students who rated instructional quality of their teacher and who gave reports on their own self-efficacy and intrinsic motivation. However, before this empirical illustration, in the following sections, I will try to describe the main concepts of the DL-MLSEM method in a (hopefully) simple and accessible style and then guide readers through its typical analytical steps—evaluating measurement models, determining reliability of variables at different level of analysis, and testing the hypothesized structural relations—with a special emphasis on the nature of examined variables (i.e., level-1 vs. level-2, climate vs. context) as well as the interpretation of their effects at each level of analysis.

6.2 Doubly Latent Multilevel Models

In educational psychology research based on multilevel designs, it is important to distinguish between level-1 (individual student-level) and level-2 (classroom or teacher-level) constructs. While the level-1 constructs are always based on responses provided by individual students, level-2 constructs can be either true level-2 measures (e.g., teacher personality, number of students in the class) or aggregated responses of students within the class, implying that level-2 constructs can be created based on level-1 reports (Marsh et al., 2012). In the construction of such level-2 variables (e.g., students' individual ratings of their teacher instructional behavior are aggregated to form class "instructional quality" climate at level-2 or students' individual academic achievement is aggregated to form class achievement), two types of error can occur. The first type refers to unreliability caused by measurement error that happens while assessing constructs either at individual or cluster level, and the second type refers to unreliability caused by sampling only a finite sample of individuals in a cluster from a potentially infinite population (Lüdtke et al., 2011).

However, the sampling error is closely tied to the nature of the level-2 constructs, that is, whether they are *formative* or *reflective* level-2 constructs. More specifically, when aggregating individual students' ratings to form the reflective construct, a group (or a class) is the referent, implying that each student directly rates some class (or teacher) level-2 characteristic (e.g., each student rates his or her teacher's enthusiasm). Since this class (or teacher) characteristic can be considered as latent construct that can be estimated based on finite numbers of students that come from potentially infinite population, ratings of all students in the class are regarded as interchangeable. In contrast, while constructing formative constructs, an individual student is the referent and the aggregation at level-2 is based on group average of individual characteristics. In that sense, level-1 ratings are no longer interchangeable since different students have different true scores that are being used to form a class average, implying that such formative aggregation corresponds to finite sampling process resulting in very small or no sampling error (Lüdtke et al., 2011). Multilevel models in which both types of errors are corrected, that is, constructs at both levels are considered latent because they are represented by multiple items (correcting for measurement error, i.e., inter-item reliability) and constructed based on responses of multiple individual raters (correcting for sample error, i.e., inter-rater agreement) are labeled as doubly latent (Lüdtke et al., 2011; Marsh et al., 2009).

6.2.1 Contextual vs. Climate Effects in Doubly Latent Multilevel Models

Related to the above-described distinction between formative and reflective level-2 constructs, contextual vs. climate effects can be differentiated. Contextual effect is defined as an effect of formative level-2 variable (based on aggregates of individual student level-1 characteristics) that exists above the effect of corresponding level-1 variable. In creating context level-2 constructs, the referent is the individual student rather than the class and ratings of different students are not interchangeable since the class average is used to represent context or classroom composition (e.g., average class achievement or class gender composition). In addition, level-1 ratings that are used to construct context variables are important and unique for their own right since they can have theoretically distinct meaning from the related level-2 context variable, implying that effects of the same variable can be quite distinct depending on the level of analysis. Therefore, to avoid bias, contextual effects should be estimated after controlling for individual student differences in the corresponding level-1 variable. A classic example of how the same level-1 variable can have distinct effects at different level of analysis is the big-fish-little-pond-effect which shows that students from high-ability schools and classes have lower academic self-concept than equally able students from low- or mixed-ability schools and classes (Marsh et al., 2007, 2008, 2009).

In contrast, climate effect is the effect of reflective level-2 construct that is also based on individual student ratings, but where each student rates the same class characteristic (e.g., teacher's displayed enthusiasm) instead of some characteristic that is specific to the student making the rating. Thus, the referent is the class rather than the individual student. Moreover, the ratings of different students within a class are theoretically interchangeable and their aggregation forms the climate construct that naturally occurs only at level 2. Consequently, differences in perceptions among students within the same class, after controlling for their shared agreement, occur at level 1 as residual climate ratings and represent a source of unreliability of the level-2 climate construct (Marsh et al., 2012). As such, residual level-1 climate ratings do not have any substantial meaning in relation to the interpretation of level-2 climate effects and, in ideal circumstances, there should be little or no systematic variance in the level-1 climate ratings after constructing the level-2 climate factor. It is important to emphasize that since the climate effect naturally occurs at classroom level (i.e., level 2), it should be primarily and dominantly studied also at the classroom level (Preacher et al., 2010, 2011). By modeling the two components of climate ratings (i.e., climate level-2 construct based on shared agreement among students within the same class and residual climate ratings at level 1) at levels they naturally occur, bias in estimating the true level-2 climate effects is reduced.

6.2.2 Measurement and Sampling Errors in Doubly Latent Multilevel Models

As in a single-level analysis, the measurement error in MSEM is controlled through introduction of latent variables defined by multiple indicators. This procedure absorbs the unreliability of imperfect indicators into their residual variances and leaves latent variables free of error, thus enabling the estimation of theoretically important relationships with greater precision. Therefore, conducting multilevel confirmatory factor analysis (MCFA) and establishing proper measurement model (i.e., confirming the hypothesized relationships between latent variable and observed indicators) oftentimes serves as an initial step upon which the structural model (i.e., hypothesized relationships between latent variables) is built.

In addition to controlling for unreliability of measurement, it is important to test whether latent variables are sufficiently conceptually and empirically differentiated to avoid issues associated with multicollinearity. This problem is especially pronounced among level-2 constructs that are based on aggregations of level-1 ratings and that tend to show higher intercorrelations when compared to individual ratings. Besides possible weaker conceptual distinctiveness between level-2 aggregations, high intercorrelations may result from partializing measurement errors from the factors as well as from the simple fact that well-functioning groups also tend to function well across indicators (Morin et al., 2021). One way to overcome this issue is specification of the bifactor model (Reise, 2012) that directly estimates global and specific factors by disaggregating the variance shared across items forming the global factor and the variance shared among items forming the specific factors (i.e., variance shared among items that are not already explained by global factor). Morin et al. (2021)

recommended that, when establishing the measurement model, a researcher should start with estimating typical correlated factor MCFA model. If there are theoretical and empirical reasons to believe that multilevel bifactor model might be appropriate (e.g., findings from previous studies or establishing high latent correlations that would suggest either conceptual redundancies or generate multicollinearity), the researcher should specify, test, and compare the multilevel bi-factor model with the correlated MCFA model, and choose the more suitable one.

Another important issue is related to measurement invariance of level-1 and level-2 constructs that are based on the individual student ratings. If the constructs have the same measurement structure with the same factor loadings at both levels (i.e., if the assumption of measurement isomorphism holds; Mehta & Neale, 2005; Morin et al., 2021), corresponding level-1 and level-2 factors can be interpreted in a similar manner. In models involving context variables, isomorphism is essential for obtaining unbiased estimates of contextual effects, while in models involving climate variables, isomorphism is mostly not relevant since researchers are rarely interested in the interpretation of the effects of level-1 residual climate ratings (Morin et al., 2021). However, imposing equality constraints, even when they are not fully supported by the theory or data, can aid to stabilization of the estimation process and lead to more accurate parameter estimates (Lüdtke et al., 2008, 2011).

It is worth mentioning that standard approaches to evaluating the goodness of model fit in both MCFA and MSEM can fail to detect the lack of fit at the higher level. Therefore, Ryu and West (2009) proposed the level-specific model fit evaluation based on partially saturated models that was found to be more informative in locating the source of lack of model fit. More specifically, they proposed that level-specific model fit indices (e.g., comparative fit index or CFI, root mean squared error of approximation or RMSEA) can be easily obtained for each level separately by saturating the model (i.e., estimating all intercorrelations between manifest variables) at the other, non-target level. For example, to obtain the fit of the measurement model at level 2, all intercorrelations between manifest variables at level 1 should be specified, and vice versa. Lastly, after establishing proper measurement model (and possibly measurement isomorphism), researchers are advised to calculate inter-item reliability of level-1 and level-2 variables by using omega (ω) coefficient of composite reliability (Geldhof et al., 2014; McDonald, 1970). It should be noted that measurement error related to constructs at level 2 tends to be smaller than the measurement error related to same level-1 constructs, thus typically resulting in greater omega (ω) coefficients at level 2.

Besides measurement errors related to inter-item agreement located at both levels of analysis, agreement between students who are rating some class characteristic (i.e., inter-rater agreement) can be considered as an additional source of error (Bliese et al., 2019; Lüdtke et al., 2008). This type of error can be estimated by calculating intraclass correlation coefficients (ICC1 and ICC2; Marsh et al., 2012). ICC1 refers to the proportion of the total variance in ratings occurring at level 2 or the average agreement between pairs of students within the same class (Bliese, 2000; Raudenbush & Bryk, 2002), while ICC2 refers to the reliability of the level-2 aggregate or a class average in relation to sampling error (Lüdtke et al., 2011). Guidelines regarding the sizes of

ICC1 and ICC2 coefficients according to which a researcher decides whether to use multilevel modeling or not, were proposed—ICC1 should be at least greater than 0.05 (Lüdtke et al., 2008), while ICC2 and omega (ω) values for variables at both levels should be larger than 0.70 (Klein & Kozlowski, 2000; Morin et al., 2014). However, doubly latent models proved to be quite robust to measurement and sampling errors and provide unbiased estimates even when the reliability is low (Marsh et al., 2012; Morin et al., 2021). Moreover, as mentioned previously, these models account for both the measurement and sampling errors, hence, their label.

6.2.3 Centering in Doubly Latent Multilevel Models

Centering (i.e., recoding scores to obtain an interpretable zero; Enders & Tofighi, 2007) is a common practice in multilevel modeling since it aids to the interpretability of multilevel models. Two types of centering should be distinguished: grand-mean centering and group-mean centering. Grand-mean centering implies subtracting the sample's mean from each score while group-mean centering implies subtracting the class (level-2 cluster) mean from each score. For constructs that contain variation at both levels of analysis (i.e., individual level-1 ratings are used to estimate level-2 construct), these two centering options can have quite different interpretational implications and, thus, their selection should be closely related to the nature of the effects under investigation. For instance, when assessing contextual effects, grand-mean centering is preferred since grand-mean ratings contain sources of variability at both levels. More specifically, level-2 effects are estimated while being controlled for their level-1 counterparts and they directly represent contextual effects or the extent to which aggregated individual characteristics add novel information when compared to simple accumulation of individual characteristics located at level 1. In contrast, when investigating climate effects, group-mean centering should be applied. By group-mean centering, level-2 aggregates become completely orthogonal to their level-1 counterparts and their effects are directly and primarily interpreted as climate effects at level 2, whereas residual level-1 ratings reflect deviations from the group average (e.g., interindividual differences in perceptions of class characteristic) and are typically not of primary interest of researchers (Morin et al., 2014, 2021).

6.3 Testing Mediation in Multilevel Structural Equation Modeling Framework

Researchers in the field of educational psychology are oftentimes interested in testing mediation hypothesis with nested data. Traditional methods for testing mediation (e.g., Baron & Kenny, 1986; MacKinnon et al., 2002) in such research designs are inappropriate due to already mentioned violation of assumption of independence of observations when having cluster data, which leads to underestimation of standard errors and possibly erroneous conclusions. Moreover, testing mediation hypotheses in standard multilevel modeling (MLM) framework is not the best choice either since MLM cannot fully separate level-1 and level-2 effects without introducing bias (Preacher et al., 2010). In overcoming these limitations, MSEM was introduced as a useful and flexible paradigm for testing indirect effects in hierarchically clustered data based on various types of multilevel research designs (Preacher, 2011; Preacher et al., 2010, 2011).

Among the most important advantages of MSEM in testing mediation hypotheses is its ability to separate variables and effects into level-1 and level-1 components to yield unbiased estimates of indirect effects at both levels of analysis. In addition, unlike MLM, MSEM allows an inclusion of mediators and outcomes that are true level-2 variables. This feature enables testing indirect effects on data from various types of research designs involving variables that have variance only at level-2 (i.e., true level-2 constructs) and variables that have variances at both levels (i.e., individual level-1 ratings are being decomposed into its level-1 and level-2 parts; Preacher et al., 2010). For example, a research design in which the predictor (X) is measured only at level 2, and the mediator (M) and outcome (Y) are measured at level-1, but their variances can be split into level-1 and level-2 components, is annotated as 2-1-1 multilevel design. Or a study in which both the mediator (M) and the outcome (Y) are measured as true level 2 variables, but the variance of the predictor (X) can be decomposed into level-1 and level-2 counterparts, has a multilevel research design annotated as 1-2-2 (for more variations in multilevel mediational designs, see Preacher et al., 2010). However, it is important to emphasize that only in the 1-1-1 multilevel design, the mediating effects can exist at both levels of analysis since the variance of all three variables in the mediational chain (i.e., predictor, mediator, and outcome) naturally exists at both levels. In all other instances where either the predictor, the mediator, or the outcome (or any of their combination) exist only at level 2, the mediating effect inherently happens exclusively at level 2. Regardless the type of the research design, when interpreting the mediating effects involving variables that exist at both levels of analysis, applied researchers should be particularly careful when ascribing substantive meaning to level-1 variables and their effects at the level 2.

Another important issue that arises when testing multilevel mediation hypothesis relates to determining statistical significance of indirect effects. Indirect effects are product terms that typically have asymmetric sampling distributions implying that standard approaches to determining confidence intervals of these estimates are not appropriate (Darlington & Hayes, 2017). In single-level analysis, this problem can be solved by calculating bias-corrected bootstrap confidence intervals (Lau & Cheung, 2012; Shrout & Bolger, 2002) or using the distribution of the product (DoP) method to obtain asymmetric confidence intervals (MacKinnon et al., 2004). Nonetheless, both options are limited in their application within multilevel mediation context (e.g., bootstrapping is not yet available for multilevel data in widely used softwares or the DoP method has not been yet extended to three-paths indirect effects). Reasonable alternatives for calculating indirect effects in multilevel mediation models include calculating Monte Carlo confidence intervals (MacKinnon et al., 2004; Preacher & Selig, 2012) or Bayesian credible intervals (Muthén & Asparouhov, 2012).

6.4 An Empirical Illustration Using Doubly Latent Multilevel Structural Equation Modeling

In the following sections, I will illustrate the application of DL-MLSEM with real multilevel data by guiding readers through steps usually taken when using this approach and with a particular emphasis on interpretation of obtained results and their understanding in relation to main DL-MLSEM concepts outlined previously. The data for this empirical illustration are taken from the study conducted by Burić and Kim (2020) and the results that will be presented here are in part already published in their paper. However, in this chapter, I will also present additional analyses and some unpublished results to provide more thorough and deeper understanding of the DL-MLSEM method.

6.4.1 Theoretical Background and Description of the Study

The aim of the Burić and Kim's study was to examine the associations between *teachers' self-efficacy* (TSE), quality of instruction they provide, and their students' motivational beliefs (i.e., self-efficacy and intrinsic motivation). A convenience sample of 94 high school Croatian teachers (gender: 86% female; years of teaching experience: M = 15.68, SD = 9.31) and 2087 students (program: 82% enrolled in a grammar-school; gender: 57% female; age: M = 16.81, SD = 0.91) participated in a cross-sectional study. On average, there were 22 students per class. Teachers provided self-reports on their TSE, while students rated instructional quality delivered by target teacher and provided self-reports on their motivational beliefs. Responses of each class of students were matched to responses of only one (target) teacher by specially created codes to avoid cross-classified data. Participation in the study was voluntary and anonymous for both teachers and their students.

TSE refers to beliefs that teachers hold about themselves regarding their capabilities to teach their subject matter and engage students in learning (Tschannen-Moran & Woolfolk Hoy, 2001). It is believed that teachers with greater TSE will invest more time in planning, be more organized and open to new ideas and methods, and persist longer in face of challenges (Tschannen-Moran & Woolfolk Hoy, 2001), hence, they will perform better. Indeed, research mostly supports this positive association between TSE and teacher performance conceptualized either as instructional quality or students' outcomes (Guo et al., 2012; Klassen & Tze, 2014; Klassen et al., 2011; Künsting et al., 2016; Midgley et al., 1989). Since the best informant regarding

TSE are teachers themselves, teachers were asked to rate their efficacy beliefs using the 10-item Teacher Self-efficacy Scale (TSE; Schwarzer et al., 1999) and 4-point response format (1 = not at all true, 4 = completely true). Sample item is "If I try hard enough, I know that I can exert a positive influence on both the personal and academic development of my students." Therefore, TSE can be considered as truly level-2 variable that has variability only at the between-teachers level.

According to contemporary definitions, *instructional quality* is viewed as a multifaceted phenomenon. For example, the Three Basic Dimensions Framework (Praetorius et al., 2018) differentiates between classroom management, cognitive activation, and supportive climate factors that are seen as critical for student learning and motivation. Classroom management refers to delivering well-structured and organized instruction with an effective student behavior management which is essential for keeping students on task most of the time. Cognitive activation implies engaging students in cognitively challenging tasks, fostering in-depth understanding of the content, and explorations of concepts and ideas, while supportive climate refers to teachers' behaviors such as constructive feedback during instruction, positive approach to students' errors and misunderstanding, as well as building caring and considerate relationships with students (Pianta & Hamre, 2009; Pianta et al., 2012; Praetorius et al., 2018).

To avoid issues related to common-method bias (Podsakoff et al., 2012) and possible self-enhancement strategies used by teachers (Schiefele & Schaffner, 2015), easily accessible and satisfactorily valid students' reports of instructional quality were used (Scherer & Gustafsson, 2015; Wagner et al., 2013). Students rated cognitive activation and classroom management dimensions on a 5-point scale (1 = strongly disagree, 5 = strongly agree), and supportive climate on a 7-point scale (1 = strongly disagree, 7 = strongly agree) by using a set of items (for more information regarding the content of the scales and their construction, see the original paper by Burić & Kim, 2020). It is important to emphasize that in all these reports, teacher was the referent, and all students were instructed to rate the same characteristic, that is, quality of instruction delivered by their teacher. In other words, students' reports of instructional quality should not be treated as reports concerning their personal, individual characteristics, but rather, conceptually, they should relate to students' "collective" class experience. Thus, students' reports of instructional quality have the status of reflective variables and effects of instructional quality are by implication the climate effects. Indeed, from the examples of items used to measure each of the three dimensions of instructional quality (i.e., "*Our* teacher gives tasks and asks questions that make *us* think," "*Our* teacher makes sure that *we* pay attention," and "*Our* teacher shows warmth to the *students*" for cognitive activation, classroom management, and supportive climate, respectively), the reflective nature of the construct becomes obvious. If the students were asked to share their personal experience (i.e., "*My* teacher makes sure that *I* pay attention"), then the effect of instructional quality could not be considered as climate effect anymore.

The quality of delivered instruction is important not only for students' academic achievement (Fauth et al., 2014; Kunter et al., 2013), but also for their *motivational beliefs* (Sakiz et al., 2012; Schiefele & Schaffner, 2015). For instance, supportive classroom climate helps to fulfill students' needs for relatedness, autonomy, and competence (Ryan & Deci, 2000), while providing meaningful and cognitively challenging tasks promotes learning goals and ensures mastery experiences, thus positively affecting students' intrinsic motivation and self-efficacy (Kunter et al., 2007). In addition, efficient classroom management keeps students focused on learning by removing disruptions and disturbances, which may enhance their interest in learning and provide opportunities for success, thus raising students' intrinsic motivation and self-efficacy levels. Students rated their self-efficacy beliefs using 8-item Self-efficacy for Learning and Achievement scale from the Motivated Strategies for Learning Questionnaire (MLSQ; Pintrich et al., 1993) and a 5-point format (1 = not at all true for me, 5 = very true of me), while the intrinsic motivation was assessed by the 4-item Intrinsic Motivation subscale of the Situational Motivation Scale (Guay et al., 2000) and a 7-point format (1 = not at all true to 7 = exactly true). It should be mentioned that students were instructed to rate their self-efficacy and intrinsic motivation in relation to the specific school subject taught by their (target) teacher to conceptually match the levels of teachers' and students' characteristics.

Burić and Kim (2020) hypothesized that TSE will be positively related to instructional quality, which will, in turn, be positively related to students' motivational beliefs. Regarding the relationship between TSE and students' motivational beliefs, they did not formulate clear expectation because of the scarcity and inconsistency of available research findings, but they did estimate these relationships in their analysis. Based on the study design described above, the hypothesized relationships imply testing the 2-1-1 mediation model in which the predictor (i.e., TSE) is strictly a level-2 variable, while the mediator (i.e., instructional quality) and outcome (i.e., students' motivational beliefs) are both level-1 and level-2 variables. Specifically, students' reports of their teacher's instructional quality and self-reports of motivational beliefs were used to create corresponding variables at both levels by partitioning their variances into level-1 and level-2 components. In addition, since one of the variables in the mediational chain (i.e., TSE) is strictly level-2 variable, possible mediation would exist and could be tested only at level 2. Finally, since the teacher was the referent in students' ratings of instructional quality, effect of instructional quality on students' motivational beliefs that occurs at level 2 is the climate effect. Nonetheless, it should be emphasized that the mediation hypothesis was not tested in the original study because of the cross-sectional and correlational nature of the research design.

6.4.2 Estimating Measurement Model and Reliability

All analyses were conducted in Mplus 8.6 (Muthén & Muthén, 1998–2017) using maximum likelihood robust estimator (MLR) and full information maximum likelihood procedures (FIML) to handle missing data (Enders, 2010). Goodness of model fit was evaluated with the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). Based on typical interpretational guidelines, CFI and TLI values greater than 0.90 and 0.95, and RMSEA values smaller than 0.06 and 0.08 indicate adequate and excellent model fit, respectively, while SRMR values smaller than 0.08 indicate good fit (Hu & Bentler, 1999; Marsh et al., 2005). In all analyses, to facilitate interpretation and reduce nonessential multicollinearity, manifest variables included in the models were standardized (M = 0.00, SD = 1.00; see also Morin et al., 2014, 2021).

Burić and Kim (2020) started their analyses with estimating ICC1 and ICC2 values to determine the amount of agreement between students who rated the quality of instruction delivered by their teacher. They found relatively high ICC1 values that ranged from 0.19 to 0.30 for composite scores and from 0.086 to 0.336 for items. In addition, ICC2 values or reliabilities of the level-2 aggregates in relation to sampling error ranged from 0.81 to 0.90 for composite scores, indicating good agreement between students while rating the same teacher for all analyzed variables (Klein & Kozlowski, 2000; Morin et al., 2014).

In the next step, they specified and tested correlated factors MCFA model. At level 2, the measurement model contained six factors (i.e., TSE, classroom management, cognitive activation, supportive climate, self-efficacy, and intrinsic motivation), while at level 1, the measurement model was consisted of five factors (classroom management, cognitive activation, supportive climate, self-efficacy, and intrinsic motivation). For all constructs that existed at both levels, the variance of their manifest indicators (i.e., items) was split into the level-1 and level-2 counterparts. At level 2, aggregated students' ratings of instructional quality stand for class "instructional quality" climate constructs, while aggregated ratings of students' self-reported self-efficacy and intrinsic motivation referred to class average self-efficacy and class average intrinsic motivation. At level-1, components of students' ratings of instructional quality represent residual ratings (i.e., individual student deviations from the group mean) and their associations with level-1 students' self-efficacy and intrinsic motivation do not have substantial meaning in relation to research hypotheses.

This model showed reasonably good fit to the data (CFI = 0.919, TLI = 0.912, RMSEA = 0.037, SRMR_W = 0.036, SRMR_B = 0.089). In addition, at level 2, TSE correlated significantly and positively with dimensions of instructional quality, but not with students' motivational beliefs, while all three dimensions of instructional quality had significant and positive latent correlations with students' motivational beliefs. At level 1, residual ratings of instructional quality also correlated positively with students' motivational beliefs. As already mentioned, residual level-1 ratings reflect deviations from the group average (i.e., interindividual differences in the perception of instructional quality provided by a teacher) and represent a source of unreliability of the level-2 climate construct (Marsh et al., 2012). Therefore, their associations with students' motivational beliefs are not of primary interest from the theoretical point of view. Even though such interindividual differences could, to a certain degree, stem from students' unique experiences of instruction, residual level-1 ratings most likely reflect perceptual differences or personal biases (Morin et al., 2021) and their relationship with students' self-efficacy and intrinsic motivation does not reflect the relationship of students' personal, individual experiences of instructional quality and their individual motivational beliefs. If we would be interested in the latter relationship, then the referent in the ratings of instructional quality should be the student instead of the teacher. For example, the item measuring supportive climate "*Our* teacher shows warmth to the *students*" should be changed to "*My* teacher shows warmth to *me*" to assess support that teacher personally give to an individual student (here, the referent is the student).

As an extension of the test of this measurement model and as additional proof of good inter-item agreement (i.e., measurement reliability) of constructs at the Level 2, omega (ω) reliability coefficients were calculated (Geldhof et al., 2014; McDonald, 1970). At level 2, omega (ω) coefficients varied between 0.922 for classroom management to 0.983 for intrinsic motivation, thus, indicating high reliability of analyzed constructs at level 2. Moreover, reliability was high even for level-1 constructs (ranging from 0.715 for classroom management to 0.918 for student self-efficacy).

Regardless the good model fit and high reliabilities, the latent correlations between different dimensions of instructional quality were high in magnitude (i.e., they ranged between 0.618 to 0.836 at level 1, and between 0.723 and 0.869 at level 2) indicating their poorer discriminant validity. Based on these results but also the theoretical relevance and conceptual distinctiveness of each of the three dimensions of instructional quality, Burić and Kim (2020) decided to continue their analyses by modeling expected relationships for each dimensions separately. However, following suggestion by Morin et al. (2021), in such cases, specifying the bi-factor model might be useful. Thus, I extended their analysis by testing the same measurement model at both levels and in which all items measuring instructional quality were specified to load on one general factor called instructional quality, while items measuring different dimensions of instructional quality were specified to load on their respective factors called classroom management, cognitive activation, and supportive climate. The multilevel bi-factor representation of instructional quality is depicted in Fig. 6.1. Other variables (i.e., TSE, students' self-efficacy, and intrinsic motivation) were specified in the same manner as in the original correlated factors model evaluated by Burić and Kim (2020). This model showed superior fit when compared to the original correlated factors model: CFI = 0.931, TLI = 0.923, RMSEA = 0.034, SRMR_W = 0.034, SRMR_B = 0.082. At level 1, the general instructional quality correlated positively with self-efficacy (r = 0.417, p < 0.001) and intrinsic motivation (r = 0.508, p < 0.001), again indicating systematic covariance between residual level-1 ratings and students' motivational beliefs. At level 2, TSE correlated positively with general instructional quality factor (r = 0.255, p = 0.032), while general instructional quality factor correlated positively to self-efficacy (r = 0.377, p < 0.001) and intrinsic motivation (r = 0.540, p < 0.001). Again, TSE was unrelated to students' self-efficacy (r = 0.019, p = 0.457) and intrinsic motivation (r = 0.001, p = 0.988). Also, the general instructional quality factor was highly reliable both at level 1 (ω = 0.929) and level 2 (ω = 0.990). Given concerns about overall measures of fit in multilevel models (Hox, 2002; Ryu & West, 2009), I additionally evaluated a partially saturated model in which, at level 1, the model was saturated by estimating intercorrelations between all manifest variables to obtain fit indices only for model at level 2. Model fit for the level-2 part was excellent: CFI = 0.977, TLI = 0.961, RMSEA = 0.024, SRMR

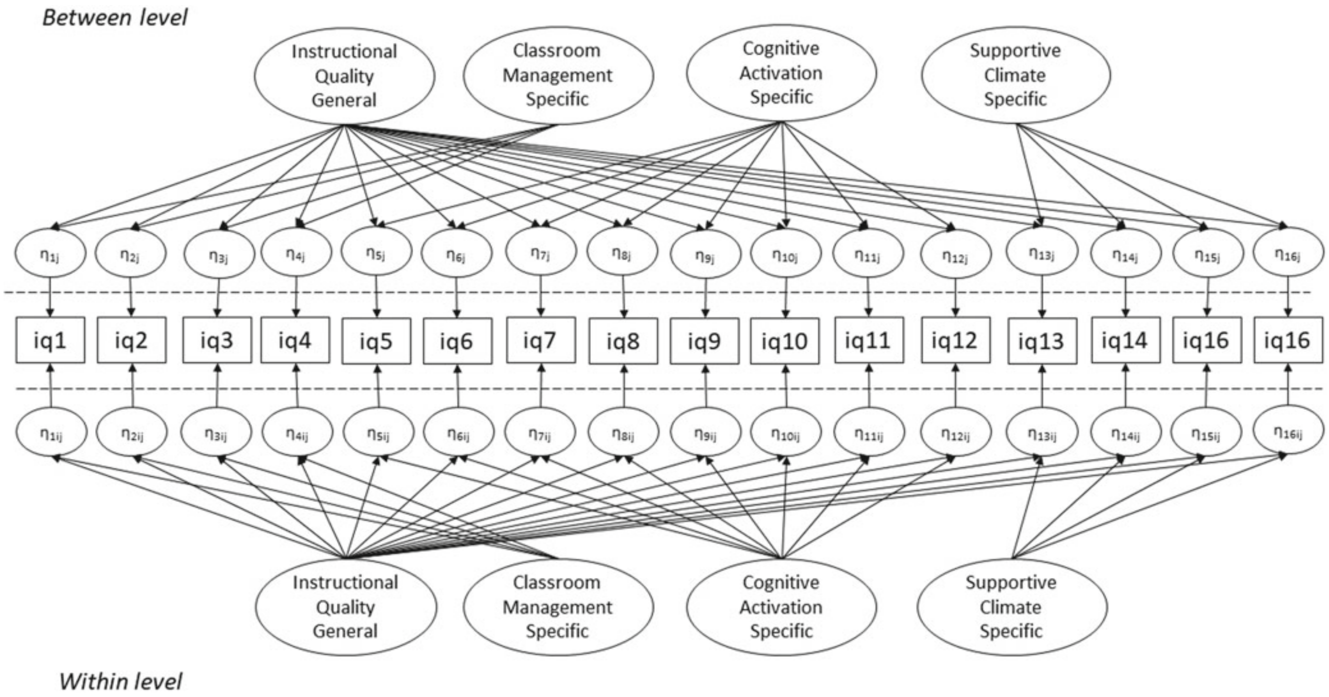


Fig. 6.1 The multilevel bi-factor measurement representation of instructional quality

= 0.083. Lastly, even though establishing the measurement isomorphism might be useful and relevant in many multilevel applications (Morin et al., 2021), considering the reflective nature of the instructional quality construct and its theoretically nonimportant effects at level 1, I opted not to test the invariance of factor loadings across levels.

6.4.3 Estimating Structural Model

In their original paper, Kim and Burić (2020) tested three structural models for each dimension of instructional quality separately. They found that TSE was positively associated with classroom management ($\beta = 0.255$ $p = 0.041$), cognitive activation ($\beta = 0.261$, $p = 0.033$), and supportive climate ($\beta = 0.234$, $p = 0.035$). In turn, classroom management, cognitive activation, and supportive climate were positively related to students' self-efficacy ($\beta = 0.361$, $p < 0.001$, $\beta = 0.312$, $p < 0.001$, and $\beta = 0.384$, $p < 0.001$, respectively) and to students' intrinsic motivation ($\beta = 0.361$, $p < 0.001$, $\beta = 0.312$, $p < 0.001$, and $\beta = 0.384$, $p < 0.001$, respectively. TSE was not related either to students' self-efficacy or to intrinsic motivation.

In the present chapter, I extended their analyses by building the bi-factor multilevel measurement model and introducing structural relationships between constructs at level 2. More precisely, at level 2, students' self-efficacy and intrinsic motivation were regressed on general factor of instructional quality and TSE, while the general factor of instructional quality was, in turn, regressed on TSE. At level 1, only latent correlations between these factors were estimated. The model showed satisfactory fit to the data: CFI = 0.931, TLI = 0.923, RMSEA = 0.034, $\text{SRMR}_\text{W} = 0.034$, $\text{SRMR}_\text{B} = 0.082$. At level 2, TSE was positively associated with general factor of instructional quality ($\beta = 0.257$, $p = 0.032$). In turn, this general factor was positively related to students' self-efficacy ($\beta = 0.371$, $p < 0.01$) and intrinsic motivation ($\beta = 0.576$, $p < 0.001$). As in the original models, TSE was not related to students' self-efficacy ($\beta = -0.008$, $p = 0.943$) or intrinsic motivation ($\beta = -0.150$, $p = 0.197$). These results are in line with those obtained in the study of Burić and Kim (2020) and lead to the same conclusions—teachers with greater TSE are rated by their students as those who deliver instruction of greater quality. In turn, average students' perceptions of greater instructional quality are related to higher class levels of self-efficacy and intrinsic motivation.

The extended analyses once again confirmed theoretical expectations and prior research findings (e.g., Dorfner et al., 2018; Guo et al., 2012; Künstig et al., 2016; Schiefele & Schaffner, 2015; Zee & Koomen, 2016) and suggested that TSE could shape students' motivational beliefs only indirectly via instructional quality. Therefore, testing the indirect effects of TSE on students' self-efficacy and intrinsic motivation seems as a theoretically justifiable approach. However, as mentioned earlier, due to the correlational and cross-sectional nature of the data, testing indirect effects would have only limited value since attributing any theoretical meaning to this possible mediating mechanism would not be completely justifiable. However, in future studies with stronger research designs (e.g., longitudinal), indirect effects could be easily obtained in Mplus by using the MODEL CONSTRAINT option and Bayes estimator or, alternatively, by implementing the Monte Carlo (MC) based parametric bootstrap approach (Preacher et al., 2010, 2011).

6.5 Conclusions

The goal of the present chapter was to provide an overview of major concepts in DL-MSEM, as an important and currently still underutilized tool in applications of multilevel modeling (Jia & Konold, 2021) as well as an illustrative example of its application. DL-MSEM allows controlling for both the measurement and sampling errors by decomposing the variance of observed variables into their level-1 and level-2 latent components to provide unbiased parameter estimates at each level of analysis (Preacher et al., 2010). In addition, due to its flexibility, it is possible to specify and test theoretically plausible but different models at each level and to obtain level specific model fit indices and measurement reliability estimates. Finally, the DL-MSEM framework proved to particularly useful for testing climate and contextual effects, as well as the mediating hypotheses (e.g., Marsh et al., 2012; Preacher et al., 2010).

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein, & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

Bliese, P. D., Maltarich, M. A., Hendricks, J. L., Hofmann, D. A., & Adler, A. B. (2019). Improving the measurement of group-level constructs by optimizing between-group differentiation. *Journal of Applied Psychology, 104*(2), 293–302. https://doi.org/10.1037/apl0000349

Bovaird, J. A., & Shaw, L. H. (2012). Multilevel structural equation modeling. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 501–518). The Guilford Press.

Burić, I., & Kim, L. E. (2020). Teacher self-efficacy, instructional quality, and student motivational beliefs: An analysis using multilevel structural equation modeling. *Learning and Instruction, 66*, 101302. https://doi.org/10.1016/j.learninstruc.2019.101302

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models.* The Guilford Press.

Dorfner, T., Förtsch, C., & Neuhaus, B. J. (2018). Effects of three basic dimensions of instructional quality on students' situational interest in sixth-grade biology instruction. *Learning and Instruction, 56*, 42–53. https://doi.org/10.1016/j.learninstruc.2018.03.001

Enders, C. K. (2010). *Applied missing data analysis.* The Guilford Press.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

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

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91. https://doi.org/10.1037/a0032138

Guay, F., Vallerand, R. J., & Blanchard, C. (2000). On the assessment of situational intrinsic and extrinsic motivation: The Situational Motivation Scale (SIMS). *Motivation and Emotion, 24*, 175–213. https://doi.org/10.1023/A:1005614228250

Guo, Y., Connor, C. M., Yang, Y., Roehrig, A. D., & Morrison, F. J. (2012). The effects of teacher qualification, teacher self-efficacy, and classroom practices on fifth graders' literacy outcomes. *The Elementary School Journal, 113*(1), 3–24. https://doi.org/10.1086/665816

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Lawrence Erlbaum Associates.

Hu, L. T., & 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. https://doi.org/10.1080/10705519909540118

Jia, Y., & Konold, T. (2021). Moving to the next level: Doubly latent multilevel mediation models with a school climate illustration. *The Journal of Experimental Education, 89*(2), 422–440. https://doi.org/10.1080/00220973.2019.1675136

Klassen, R. M., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review, 12*, 59–76. https://doi.org/10.1016/j.edurev.2014.06.001

Klassen, R. M., Tze, V. M. C., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998–2009: Signs of progress or unfulfilled promise? *Educational Psychology Review, 23*, 21–43. https://doi.org/10.1007/s10648-010-9141-8

Klein, K. J., & Kozlowski, S. W. J. (Eds.). (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions.* Jossey-Bass.

Künsting, J., Neuber, V., & Lipowsky, F. (2016). Teacher self-efficacy as a long-term predictor of instructional quality in the classroom. *European Journal of Psychology of Education, 31*, 299–322. https://doi.org/10.1007/s10212-015-0272-7

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*, 494–509. https://doi.org/10.1016/j.learninstruc.2007.09.002

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820. https://doi.org/10.1037/a0032583

Lau, R. S., & Cheung, G. W. (2012). Estimating and comparing specific mediation effects in complex latent variable models. *Organizational Research Methods, 15*(1), 3–16. https://doi.org/10.1177/1094428110391673

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467. https://doi.org/10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203–229. https://doi.org/10.1037/a0012869

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*(1), 83–104. https://doi.org/10.1037/1082-989X.7.1.83

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*(1), 99–128. https://doi.org/10.1207/s15327906mbr3901_4

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Lawrence Erlbaum Associates Publishers.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764–802. https://doi.org/10.1080/00273170903333665

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*(3), 319–350. https://doi.org/10.1007/s10648-008-9075-6

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal, 44*(3), 631–669. https://doi.org/10.3102/0002831207306728

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*(1), 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Change in teacher efficacy and student self- and task-related beliefs in mathematics during the transition to junior high school. *Journal of Educational Psychology, 81*, 247–258. https://doi.org/10.1037/0022-0663.81.2.247

Morin, A. J., Blais, A. R., & Chénard-Poirier, L. A. (2021). Doubly latent multilevel procedures for organizational assessment and prediction. *Journal of Business and Psychology*, 1–26. https://doi.org/10.1007/s10869-021-09736-5

Morin, A. J., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education, 82*(2), 143–167. https://doi.org/10.1080/00220973.2013.769412

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In *Handbook of advanced multilevel analysis* (pp. 23–48). Routledge.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*, 109–119. https://doi.org/10.3102/0013189X09332374

Pianta, R. C., Hamre, B. K., & Mintz, S. L. (2012). *The CLASS-secondary manual*. University of Virginia.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement, 53*, 801–813. https://doi.org/10.1177/0013164493053003024

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

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/1007/s11858-018-0918-4

Preacher, K. J. (2011). Multilevel SEM strategies for evaluating mediation in three-level data. *Multivariate Behavioral Research, 46*(4), 691–731. https://doi.org/10.1080/00273171.2011.589280

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*(2), 77–98. https://doi.org/10.1080/19312458.2012.679848

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling, 18*(2), 161–182. https://doi.org/10.1080/10705511.2011.557329

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods, 21*, 189–205. https://doi.org/10.1037/met0000052

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209–233. https://doi.org/10.1037/a0020141

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16*(4), 583–601. https://doi.org/10.1080/10705510903203466

Sakiz, G., Pape, S. J., & Hoy, A. W. (2012). Does perceived teacher affective support matter for middle school students in mathematics classrooms? *Journal of School Psychology, 50*, 235–255. https://doi.org/10.1016/j.jsp.2011.10.005

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 factor structural equation modeling. *Frontiers in Psychology, 6*, 1–15. https://doi.org/10.3389/fpsyg.2015.01550

Schiefele, U., & Schaffner, E. (2015). Teacher interests, mastery goals, and self-efficacy as predictors of instructional practices and student motivation. Contemporary *Educational Psychology, 42*, 159–171. https://doi.org/10.1016/j.cedpsych.2015.06.005

Schwarzer, R., Schmitz, G. S., & Daytner, G. T. (1999). *The Teacher Self-Efficacy scale* [On-line publication]. http://www.fu-berlin.de/gesund/skalen/t_se.htm

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*(4), 422–445. https://doi.org/10.1037/1082-989X.7.4.422

Silva, B. C., Bosancianu, C. M., & Littvay, L. (2019). *Multilevel structural equation modeling*. Sage Publications.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage publications.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*, 783–805. https://doi.org/10.1016/s0742-051x(01)00036-1

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003

Zee, M., & Koomen, H. M. Y. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being. *Review of Educational Research, 86*, 981–1015. https://doi.org/10.3102/0034654315626801

Zyphur, M. J., Zhang, Z., Preacher, K. J., & Bird, L. J. (2019). Moderated mediation in multilevel structural equation models: Decomposing effects of race on math achievement within versus between high schools in the United States. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 473–494). American Psychological Association.

Irena Burić, Ph.D. is an Associate Professor at the Department of Psychology of University of Zadar, Croatia. She teaches courses in educational psychology and basic and advanced statistics at undergraduate, graduate, and postgraduate levels. She has attended numerous postgraduate courses, workshops, and summer schools in structural equation modeling, multilevel analysis, longitudinal analysis, and item response theory. Her research interests are directed toward investigating teacher personality, emotion, and emotion regulation and their associations with teacher well-being and performance. She is a member of the editorial board in the Personality and Individual Differences and Psychological Test Adaptation and Development journals. She serves as a reviewer in many distinguished journals (e.g., Teaching and Teacher Education, Personality and Individual Differences, Motivation and Emotion, Learning and Instruction).

Part III
Methodology for Multilevel Modeling

Chapter 7
Analyzing Large-Scale Assessment Data with Multilevel Analyses: Demonstration Using the Programme for International Student Assessment (PISA) 2018 Data

Julie Lorah

Abstract The present chapter provides a tutorial with example demonstration designed to guide the reader through the complexities and unique challenges associated with analyzing large-scale assessment data through multilevel modeling techniques. The reader will learn how to use relevant literature to specify an appropriate model considering effects at various levels; how to address large-scale data complexities including sampling weights and plausible values; and how to estimate and interpret the results from such a model. The chapter will provide a conceptual background and short description of each of these topics, followed by an example demonstration using Programme for International Student Assessment (PISA) 2018 data. The demonstration will be used to provide a concrete example for analysis, including the process of model specification, estimation, and interpretation. In addition, annotated R syntax and R output associated with aspects of modeling will be provided so that readers will easily be able to conduct their own multilevel analyses with PISA data in order to answer their own research questions.

Keywords Large-scale assessment · PISA · Model specification · Estimation · R · Plausible value

7.1 Introduction

Data from international large-scale assessments can be used by researchers to answer various questions about students around the world. The present study will demonstrate use and analysis of one of these assessments, the Programme for International Student Assessment (PISA), which comprises a generalizable sample of 15-year-old students from several countries and is administered every three years. The goal of PISA is to assess students near the end of their compulsory education to understand the extent to which the students are prepared to participate in society (OECD, 2009). In addition to formal assessments in math, science, and reading, PISA collects survey

J. Lorah (✉)
Indiana University, Bloomington, IN, USA
e-mail: jlorah@iu.edu

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_7

data from the students, their parents, and their schools. Although data from PISA represents a huge opportunity for researchers, it may not be simple to analyze the data due to the complex nature of data collection.

There are many models available to analyze assessment and survey data. However, due to the nested nature of the data from large-scale assessments, and the multi-stage sampling methods often used, multilevel modeling represents a particularly useful strategy for analysis of this data. A multilevel model explicitly includes group membership information, such as school membership. Ignoring this type of nesting is not advised, as it has been shown to cause an increase in Type I error rates (Snijders & Bosker, 2012). Additionally, use of a multilevel model easily allows for inclusion of contextual effects, such as school-level variables.

Previous research has used multilevel modeling techniques to analyze PISA data. For example, Ferrao et al. (2017) used a multilevel binary logistic regression to model the outcome of student grade repetition. Several predictors were included in the model at both the student and school level. In another study (Ertem, 2021), PISA 2018 was used to examine reading literacy among students in Turkey. Student-level predictors, such as disciplinary climate and enjoyment of reading were included, along with school-level variables, such as the proportion of parents involved in school.

The purpose of the present study is to provide a tutorial for researchers interested in analyzing PISA 2018 data using a multilevel model. Although the present study focuses on PISA 2018, similar methods can be used for analysis of other large-scale assessment data.

7.2 PISA 2018 Data

The target population for PISA 2018 is 15-year-old students attending educational institutions (OECD, 2021a). The sampling plan involved a two-stage stratified sample where schools within each country were stratified and then sampled with probability proportional to size; and then students within selected schools were sampled. The exception to this was Russia where a modified sampling plan was used whereby a three-stage sampling plan with the additional factor of geographical area was added. The stratification variables and number of strata used to create strata for schools were different for each country. To select students, a random sample of 35 or 42 students from the eligible population was selected with equal probability within each selected school. For schools with less than the pre-determined sample size (i.e. 35 or 42), all students were selected. A total of 80 countries participated (OECD, 2021a).

7.3 Complex Survey Design

Data from large-scale assessments is typically not collected through a simple random sample. Therefore, aspects of the complex sampling plan must be accounted for during the analysis of this data. In addition, many large-scale assessments provide several plausible values to measure achievement, rather than a single value. These plausible values require accounting for during analysis as well. Each of these aspects related to complex survey design in the PISA 2018 data is described in more detail below.

7.3.1 Sampling Weights

The students selected for the PISA 2018 assessment were not selected with equal probability. Since the probability of selection for each student differs, this must be accounted for when analyzing the data in order to avoid bias in parameter estimates (Lorah, 2019). Sampling weights, also referred to as survey weights, represent the inverse of the probability of selection for the given observation (Snijders & Bosker, 2012). In the present analysis, sampling weights are automatically incorporated into the analyses of descriptive statistics through the "intsvy" package within R, and into the estimation of the multilevel model through the "WeMix" package in R.

7.3.2 Plausible Values

Rather than measure achievement as a single point estimate, large-scale assessments often include several plausible values for student achievement measures. For example, in PISA 2018 math achievement is measured by a set of 10 plausible values for each student. In other words, there are 10 variables included in the dataset to measure math achievement. For each student, each of those 10 plausible values will have a slightly different value. Each plausible value represents a random draw from a distribution of possible values (i.e. the posterior distribution) for that student's achievement and is typically a function of both the student's item responses as well as their survey question responses (Martin & Mullis, 2012). Plausible values are used in place of point estimates in order to model uncertainty in score estimation (Lorah, 2019). In order to conduct an analysis with plausible values, the model or statistic will be estimated separately for each plausible value and then the estimates will be combined. Please note that it is never advised to take the average of plausible values for use in analysis (Rogers & Stoeckel, 2008).

For example, with PISA 2018 data, a model using math achievement as an outcome variable would be estimated 10 times: one time for each of the 10 plausible values. The estimates from those 10 models can then be combined to create parameter estimates and standard errors which represent the final values for reporting.

The procedure for combining the multiple values goes as follows. To compute the point estimate for a given statistic, just compute the statistic once for each plausible value, and then take the average. To compute the standard error, the following formula may be used:

$$\text{SE} = \text{sampling variance} + (1 + 1/\text{M}) * \text{Var}(t1, \ldots, t\text{M})$$

Where M represents the number of plausible values; for example for PISA 2018, M = 10. The sampling variance can be computed by taking the mean of all M standard error estimates; and the *t* represents the M estimates for which the standard error is being computed (Martin & Mullis, 2012). For example, imagine that a linear regression model is estimated using PISA 2018 data with math achievement used as an outcome and we are interested in finding the correct intercept estimate and its standard error. First, the model will be estimated 10 times; one for each math plausible value. The first model will use the first math plausible value as an outcome and the intercept estimate from this first model represents $t1$; the intercept from the second model represents $t2$, etc. The sampling variance is computed as the variance of all 10 intercept standard error estimates. Computations associated with plausible values are demonstrated in the example analysis below.

7.3.3 Replicate Weights

One additional factor that may need to be considered is the use of replicate weights. These weights allow for correct estimation of standard error estimates. However, when the nesting structure in the model is consistent with that of the sampling plan, these are not needed. Specifically, when the grouping variables in the model correspond to the stages in the sampling design, use of replicate weights is not necessary (Snijders & Bosker, 2012). In the present analysis, the sampling design initially samples schools, and then students. Given that both these levels are explicitly accounted for in the multilevel model, the present analysis will not use replicate weights.

7.4 Specifying a Multilevel Model

Depending on the research question and available literature, model specification may have elements of more exploratory approaches and/or more confirmatory approaches. When relevant theories are available, these should be used to inform the predictors included in a model, including fixed effects, random effects (i.e. nesting structure), and interaction effects. When relevant theories are less available, researchers may choose to take a more exploratory approach and specify a model based on their research question, subject knowledge, and previous empirical studies on the topic. When appropriate, this more exploratory approach can be conducted with a procedure described by Snijders and Bosker (2012) whereby the researcher sequentially tests possible fixed effects at level one; followed by fixed effects at level two; followed by random effects such as a third level of nesting or random slopes; and lastly, non-linear predictors, such as interaction terms (see Snijders & Bosker, 2012, chapter 6 for more details). Regardless of the method chosen to arrive at a final model specification, the researcher should be clear about this procedure, including any preliminary models that were estimated, when describing the research methods.

Another decision that is needed in order to specify a multilevel model is whether each categorical predictor variable will be modeled as a random effect or a fixed effect. Generally, any variable where there is interest in the specific categories should be modeled as a fixed effect, whereas any variable where interest lays in the distribution of the outcome across all categories can be modeled as a random effect. In addition, researchers recommend that the data should include at least about 10 groups in order to model group as a random effect (Snijders & Bosker, 2012).

When specifying a multilevel model, researchers have the flexibility to model predictors measured directly at level one, directly at level two, or aggregate measures. For example, individual student SES could be included as well as the aggregate measure of average SES within a school. A model including both the individual and group-aggregated variable is called a contextual model (Ma et al., 2008) and can be a useful model in educational research.

7.5 Example Analysis

The following sections describe an applied analysis using multilevel modeling with PISA 2018 data. Note that this analysis is provided for demonstration purposes as opposed to specific substantive claims. All syntax to complete the analysis is provided in the appendix. Since the data is available to download from the OECD, readers may choose to use this syntax to replicate the analysis on their own.

7.6 Obtaining the Data

The PISA 2018 data is freely available online (https://www.oecd.org/pisa/data/). This tutorial will use the "student questionnaire data file" and the "school questionnaire data file." The code book is also available from this source and the OECD provides an overview of the data and important aspects of analysis for researchers (OECD, 2021b). The data is available in two formats: SAS data files and SPSS data files. To open the data in R (R Core Team, 2021), save the data files in the working directory folder, and then use the "intsvy" package (Caro & Biecek, 2017) to load the data. The "pisa.select.merge" function from this package allows the user to specify multiple files, variables, and/or countries. If multiple files are selected, the function automatically merges the datasets from each file. The present analysis uses both the school file and student file. Multiple variables from each file are selected and one country, the United States, is used for this analysis. This function also automatically selects all achievement and weighting variables, in addition to the variables specified by the user. In order to verify the coding for each variable and the response categories in the dataset, the “pisa.var.label” function from this “intsvy” package may be used. This function will save a text file with a list of variables from the files selected which provides a succinct summary and description of the dataset.

Sections (1) and (2) of the R syntax in the appendix demonstrate these steps. In Section (1), the two packages used for analysis are imported: instvy is used to import the data and compute descriptive statistics, while WeMix is used to estimate the multilevel models with sampling weights. In addition, it is important to set an appropriate working directory, which is where the data files should be saved. In Section (2) of the R syntax, a file with the variable labels is saved (i.e. codebook) and the data is merged and imported into R. The file with variable labels will be saved into the working directly by using the pisa.var.label function. Within the function, the user specifies which data files are being used: in the present study the school and student file are being used and so their file names are specified within the syntax. By running this function, the user will benefit from the creation of a file (saved in the working directly) that concisely summarizes the variable names and labels describing what the variable represents, similar to a code book. The pisa.select.merge function is then used to merge multiple dataset and then import those datasets into R. In this example, two datasets, the student file and school file, are used. The user then has the ability to specify the specific variables to import from each file, as well as the specific countries. After the data is merged and imported, it is stored in a data frame called “mydata” in the present example.

7.7 Model Specification

The model specification procedure for the present analysis was a bit less rigorous compared with a substantive study since the primary purpose of this analysis is to provide a tutorial. The nesting structure for these models is represented by students nested within schools. As one might imagine, schools represent a source of non-independence in the data and should be included in analysis. Since the primary interest is in the distribution of schools, rather than individual schools, the school is modeled as a random effect in all models considered. The first two models (M1 and M2) consider a math achievement outcome variable, and the second two models (M3 and M4) consider an outcome variable measuring the student’s mastery achievement orientation. Understanding the relationship between various predictors and these outcomes could assist researchers and educators in their understanding of these constructs. Ideally, a set of predictor variables, including control variables will be selected based on a well-defined research question, as well as the available literature. In the present analysis, predictors from both level one and level two are selected, including a measure of socioeconomic status, student gender, a measure of student fear of failure, and the number of students in the school. The first three predictors are measured at the individual student level and the fourth predictor is measured at the school level.

7.8 Analysis

Prior to estimating models, the data is prepared and cleaned and descriptive statistics are computed. These steps are demonstrated in Sections (3) and (4) of the R syntax. In Section (3), two new variables are created. First, a new variable called "SCHSIZE_TH" is created by dividing the school size variable by 1000. This new variable is now in units of "thousands of students" compared with the original variable for which the units are "students." This transformation is helpful for both interpretation and potentially for model estimation/convergence, since when variables are on very different scales, this can sometimes lead to convergence issues. The second new variable, "Male" is simply a new name for the original variable. Since interpretation is enhanced when the variable is named in the direction of the effect, the name "Male" is more descriptive than a variable name such as "Gender." Note that the variable could have been coded in the opposite direction and named "Female"; the choice is arbitrary.

In Section (4) the computation of relevant descriptive statistics is demonstrated. First, the "pisa.mean.pv" function is used to compute the mean value for plausible values. Note that the function takes the variable label "MATH" to indicate all 10 plausible values (PV1MATH, PV2MATH,…, PV10MATH). This function provides the mean and standard deviation for math plausible values by automatically combining results from all 10 plausible values appropriately. The function also automatically incorporates complex survey design elements, such as sampling weights. For variables that don't involve plausible values, the "pisa.mean" function is used, which also automatically incorporates the complex survey design. In addition, the "pisa.table" function is used to provide the frequency distribution for categorical variables and the "pisa.ben.pv" function can provide the percentages at each proficiency level for variables using plausible values. Subgroup means can be computed with these functions by including the "by=" argument. For example, the subgroup means on the math plausible values are computed for males and females, again using the "pisa.mean.pv" function. Lastly, correlation coefficients are computed. Although the "pisa.rho" function can be used to compute correlations involving complex survey design, the plausible values need to be combined by the user to arrive at final correlation estimates. This is shown in Table 7.1 where correlations across three math plausible values are combined. Note that only three plausible values are used for simplicity of demonstration purposes, but in general analyses should incorporate all 10 plausible values.

For each of the two outcome measures (math achievement and mastery goal orientation), two multilevel models are run. First, the empty, or unconditional model is run to assess the degree of nesting (Snijders & Bosker, 2012):

$$\mathrm{Y}_{ij} = \gamma_{00} + \mathrm{U}_{0j} + \mathrm{R}_{ij}$$

$$\mathrm{Var}\left(\mathrm{R}_{ij}\right) = \sigma^2$$

Table 7.1 Descriptive statistics

				Correlations					
Variable	N	M	SD	MATH	MASTGOAL	ESCS	MALE	GFOFAIL	SCHSIZE_TH
MATH	4838	478.240	92.140	1.000					
MASTGOAL	4608	0.291	0.990	−0.037	1.000				
ESCS	4767	0.106	1.015	0.398	0.044	1.000			
MALE	4838	1.511	0.500	0.062	−0.180	0.018	1.000		
GFOFAIL	4626	0.168	1.052	0.169	0.030	0.131	−0.198	1.000	
SCHSIZE_TH	4279	1.490	0.917	0.119	0.011	0.066	0.003	0.051	1.000

	PV1MATH	PV2MATH	PV3MATH	Average Correlation
MASTGOAL	−0.030	−0.040	−0.040	−0.037
ESCS	0.396	0.394	0.404	0.398
MALE	0.054	0.066	0.066	0.062
GFOFAIL	0.171	0.166	0.171	0.169
SCHSIZE_TH	0.135	0.105	0.118	0.119

	N	Percentage	Mean Math Score	SD, Math Scores
Female	2376	48.92	473.85	88.78
Male	2462	51.08	482.46	95.05

Note The first portion of this table displays the final descriptive statistics, appropriate for reporting. Statistics associated with plausible values have been combined appropriately, as described in the text. The second portion demonstrates the computation of the combined estimates for each correlation coefficient associated with the math plausible values. The third portion shows the mean math plausible value, disaggregated by group membership. All statistics were computed using functions from the intsvy package, automatically accounting for the complex nature of the data, such as inclusion of sampling weights

$$\mathrm{Var}\big(\mathrm{U}_{0j}\big) = \tau^2$$

Where Y_{ij} represents the continuous outcome, γ_{00} represents the intercept, U_{0j} represents level-two residuals, R_{ij} represents level-one residuals, i represents individual students (level-one), and j represents schools (level-two). The empty model is used to compute the intraclass correlation coefficient (ICC):

$$\mathrm{ICC} = \tau^2/\big(\tau^2 + \sigma^2\big)$$

The ICC can be interpreted as the proportion of variation at level two for the given outcome measure. In the present example, the ICC will be interpreted as the proportion of variability in math achievement (or mastery goal orientation) at the school level.

Second, the full model including all predictors will be estimated:

$$\begin{aligned}\mathrm{Y}_{ij} = {} & \gamma_{00} + \gamma_{01} * \mathrm{ESCS}_{ij} + \gamma_{02} * \mathrm{Male}_{ij} + \gamma_{03} * \mathrm{GFOFAIL}_{ij} \\ & + \gamma_{10} * \mathrm{SCHSIZE\_TH}_{j} + \mathrm{U}_{0j} + \mathrm{R}_{ij}\end{aligned}$$

$$\mathrm{Var}\big(\mathrm{R}_{ij}\big) = \sigma^2$$

$$\mathrm{Var}\big(\mathrm{U}_{0j}\big) = \tau^2$$

These two models (the empty model and the full model) will each be estimated for both outcome measures (math and mastery goals), resulting in four total models. The level one predictors include ESCS (a continuous measure of SES); Male (a binary measure of gender where female is coded 1 and male is coded 2); GFOFAIL which measures general fear of failure; and SCHSIZE_TH which is a measure of the number of students in the school divided by 1000.

Section (5) of the R syntax demonstrates estimation of the multilevel models. In order to estimate these four models, the "mix" function from the "WeMix" package (Bailey et al., 2021) is used. This function is convenient as the syntax mirrors that of another popular multilevel modeling package, lme4 (Bates et al., 2015). The "WeMix" package allows the user to include sampling weights at each level of the model. Within Section (5) of the R syntax, the first set of models involves the math outcome, which makes use of plausible values. The second set of models involves the MASTGOAL outcome, which does not involve any plausible values. Within the first set of models examining the math outcome, the first step is to listwise delete the analytic dataset. In order to do so, only the variables used in the present set of models are considered. Then, six models are estimated with the math outcome. The first three (M1a, M1b, and M1c) are the empty model, which will be used to compute the ICC. Three models are estimated instead of one single model in order to account for the plausible values. These three models are identical, except that M1a uses the first plausible value as outcome (PV1MATH), M1b uses the second plausible value as outcome (PV2MATH), and M1c uses the third plausible values (PV3MATH) as outcome. After these three models are estimated, the coefficients can be combined accordingly. Note that only three plausible values are used to simplify demonstration; however, for applied purposes all 10 plausible values should be used, in which case 10 empty models would be estimated. In these three models, the variable CNTSCHID represents the school ID which is the nesting variable, or the level 2 variable. The weights specified are the individual and school-level weights. The next set of three models (M2a, M2b, and M2c) are similar to the first three, but additionally incorporate relevant predictors.

Models involving the MASTGOAL outcome are specified similarly. Again, listwise deletion is performed on the analytic dataset. Following this one empty model (M3) and one full model with all predictors (M4) are estimated. Since no plausible values are involved, multiple iterations of the same model do not need to be specified.

Section (6) of the R syntax demonstrates the appropriate procedure for combining results based on plausible values. The intercept estimate from the empty model (based on M1a, M1b, and M1c) is combined for demonstration purpose; however, in an applied analysis all relevant parameter estimates should be combined analogously to arrive at final model estimates appropriate for reporting. In order to compute the final intercept estimate, first, the point estimate for this coefficient is computed as the average of the individual point estimates associated with each plausible value, respectively. In this case, the intercept estimate from M1a (with PV1MATH outcome), from M1b (with PV2MATH outcome), and from M1c (with PV3MATH outcome) are all averaged to arrive at the final intercept estimate appropriate for reporting. Note that in an applied analysis, all 10 plausible values would be used. In order to compute the appropriate standard error, the formula provided above is used. First, the sampling variance is computed as the average of the standard error for the intercept from M1a, from M1b, and from M1c. Next, M is specified to equal 3, since there are three plausible values used (for applied analyses, all 10 should be used). The last term in the equation computes the variance associated with the set of intercept estimates. This procedure for computing the point estimates and its associated standard error should be then used for all reported results. Although the specific computations for doing so for the rest of the coefficients are not demonstrated in Section (6) of the R syntax, the procedure is analogous, and the results of doing so can be viewed in Table 7.2.

7.9 Results

7.9.1 Descriptive Statistics

Descriptive statistics results can be found in Table 7.1, including each variable's mean and standard deviation, and all correlations. The results show that MASTGOAL, ESCS, and GFOFAIL appear to be approximately standardized, with mean close to zero and standard deviation close to one for each variable. The sample

Table 7.2 Model results

		PV1MATH		PV2MATH		PV3MATH		Combined	
		Coef	SE	Coef	SE	Coef	SE	Coef	SE
M1	Intercept	474.000	5.360	476.894	5.399	474.530	5.114	475.141	8.456
	School Var	1784.000		1764.000		1795.000		1781.000	
	Residual Var	6371.000		6394.000		6338.000		6367.667	
M2	Intercept	444.730	7.434	445.416	7.678	442.265	6.952	444.137	11.016
	ESCS	25.174	1.669	25.574	1.753	25.575	1.676	25.441	1.771
	MALE	13.404	2.852	16.055	2.711	15.753	2.875	15.071	5.621
	SCHSIZE_TH	7.734	3.782	4.967	3.799	6.933	3.402	6.545	6.364
	GFOFAIL	7.541	1.276	7.452	1.283	7.509	1.223	7.501	1.263
	School Var	992.600		1011.000		1001.000		1001.533	
	Residual Var	5821.800		5816.000		5761.000		5799.600	

Note Both models are based on sample size of 4074 students nested within 143 schools

is comprised of approximately 51% male students and 49% female students. The average school size is reported as about 1.49 which represents how many thousand students are in the school. In other words, the average school size is 1,490. Correlations indicate that math achievement is not strongly correlated with any predictor. The largest correlation is with ESCS at about 0.4. Generally the predictors are not highly correlated with each other, except the correlation of about −0.2 between Male and GFOFAIL indicating males are less likely to indicate a fear of failure than females. The “intsvy” package also allows simple computation of summary statistics by group. For example, the mean math score by gender is included in the final part of the table. Note that since these statistics were all computed with the appropriate function from the “intsvy” package, correct treatment of complex survey data, such as inclusion of sampling weights, is automatically produced.

With the “intsvy” package, these descriptive statistics are straightforward to compute with the exception of correlations for plausible values. In order to compute these correlations, the procedure for combining estimates with plausible values was used. Specifically, since the point estimate for a statistic is simply the mean of the statistic computed once for each plausible value, the correlations for this data are simply the mean of the correlations computed with each plausible value (Martin & Mullis, 2012). For example, the correlation between Math and MASTGOAL is reported as −0.03667. This is computed as the mean of three correlations; −0.03 is the correlation between PV1MATH and MASTGOAL; −0.04 is the correlation between PV2MATH and MASTGOAL; and −0.04 is the correlation between PV3MATH and MASTGOAL. Each correlation between MATH and any other variable can be computed analogously. Note that only 3 plausible values (PV1MATH, PV2MATH, and PV3MATH) were used for the present demonstration for clarity of presentation; however, when analyzing PISA data, all plausible values should be used; PISA 2018 includes 10 plausible values for each achievement variable. It is straightforward to include all 10 plausible values, rather than this subset of 3 plausible values, in a final analysis by combining estimates from multiple models analogously.

7.9.2 Math Achievement Outcome

Two models with the math achievement outcome were estimated: the empty model (M1) and the full model (M2). However, since math achievement is measured with several plausible values in PISA 2018, both M1 and M2 need to be estimated multiple times—once for each plausible value. Although all 10 plausible values should be used for applied analyses, the present demonstration shows the process with just the first 3 plausible values for clarity of presentation. The method for combining these models is described above and demonstrated here.

Table 7.2 shows the model results. All models are based on an analytic sample of 4074 students nested within 143 schools. The results are listed for six total models: PV1MATH was the outcome for M1a and M2a; PV2MATH was the outcome for M1b and M2b; and PV3MATH was the outcome for M1c and M2c. The last two columns show the combined results which are what would be reported in an applied analysis. To combine the results from multiple models, the coefficient estimates may simply be averaged. This is demonstrated in Table 7.2, as well as in the R syntax. For example, to find the intercept estimate from the empty model, the average of all three intercept estimates gives a combined value of 475.1413. The column displaying the combined coefficients are all computed analogously as the average of the given coefficient from a model with PV1MATH, PV2MATH, and then PV3MATH as outcome, respectively. The formula to compute the standard error is given above as well as in the R syntax and used in the table to arrive at the combined estimates for the standard error values.

Results from M1 provide the random effects which can be used to compute the ICC. The ICC should be computed based on the final (combined) variance component estimates given for M1. This results in an ICC of about 0.22 according to the ICC formula given above. This indicates that about 22% of the variance is math achievement is found at the school level. For M2, a t value can be computed by dividing each fixed effect coefficient by its respective standard error. Doing so indicates that all four predictors are significant. After controlling for associated covariates, results indicate that: one standard deviation increase in ESCS is related to about 25 points increase in math scores; males score on average about 15 points higher than females; increasing the school size by 1000 students is related to about 6.5 points increase in student math achievement; and about 1 standard deviation increase in fear of failure is related to about 7.5 points increase in math achievement.

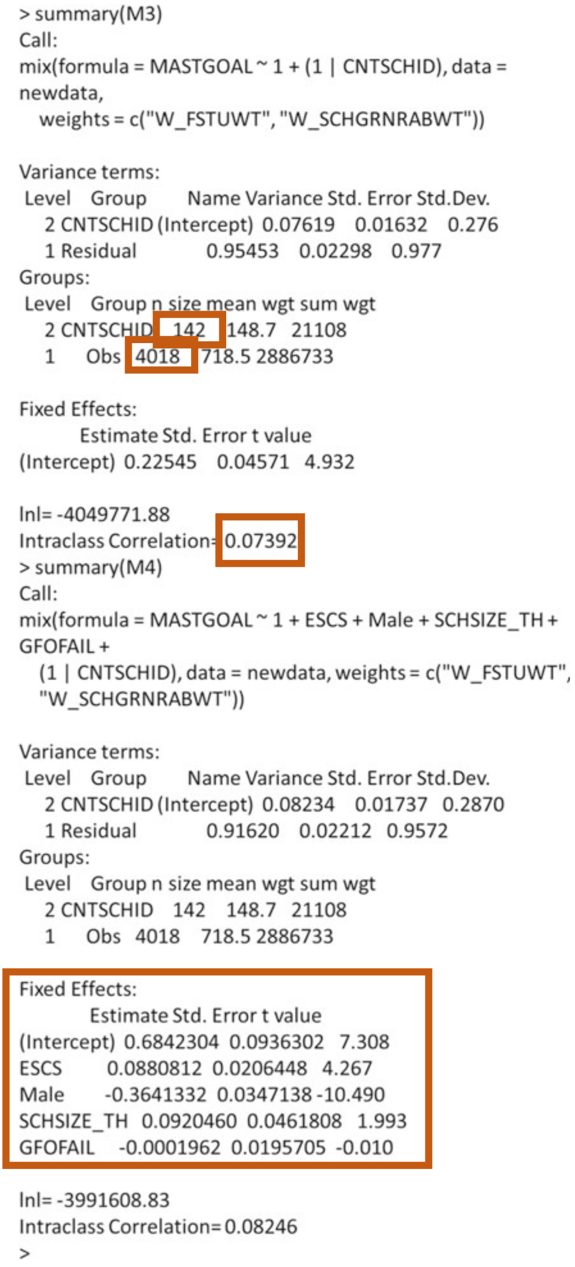
7.9.3 MASTGOAL Outcome

Two models with the mastery goal orientation outcome were estimated in order to demonstrate models not using plausible values: the empty model (M3) and the full model (M4) results are provided in the annotated output in Fig. 7.1 and in Table 7.3. The annotated output in Fig. 7.1 shows the output directly from R. The relevant quantities for interpretation are indicated in Fig. 7.1 with a box around those numbers. In order to view the results of these two models, the "summary" function is used. Table 7.3 displays a succinct summary of the results from M3 and M4, similar to what might be included in a final research report.

After listwise deletion, the final analytic sample for these two models was 4018 students nested within 142 schools. The primary result from the empty model is the ICC, which can be used as a measure of effect size for the random effect (Lorah, 2018). The ICC is about 0.07 indicating that about 7% of the variance in student mastery goal orientation is at the school level. Alternatively, the ICC may be interpreted as the correlation between two randomly drawn students from one randomly drawn school (Snijders & Bosker, 2012). This ICC value seems reasonable since schools may have different cultures and populations that relate to this outcome. However, this ICC value of 0.07 indicates that most of the variance in mastery goal orientation is not at the school level; but rather than there is a large amount of individual difference in terms of mastery goal orientation. Note that this value is a bit less than the ICC for math achievement which is about 0.22. It seems reasonable that there is more variation at the school level for math achievement than for mastery goal orientation.

The results from M4 indicate that all predictors, except GFOFAIL are significantly related to mastery goal orientation. In addition to interpreting significance, it is important to interpret each specific regression coefficient, or provide an interpretable effect size measure. Since the units for each of these predictors is already intuitively interpretable, the unstandardized coefficients given from the output are appropriate to interpret. For example, according to the descriptive statistic results, the standard deviation of both MASTGOAL and ESCS is about one. To interpret the coefficient for this predictor, we can state that for every one standard deviation increase in ESCS, we expect the student's mastery goal orientation to increase by about 0.09 standard deviation units, after controlling for associated covariates. Analogously, we expect males to score about 0.36 standard deviations lower on mastery goal orientation compared with females; and we expect for every additional 1000 students in a school's population, the student would increase on average about 0.09 standard deviation units in mastery achievement orientation, after controlling for associated covariates. There is no evidence of a relationship between the student's general fear of failure and mastery goal orientation, after controlling for associated covariates. Note that we don't necessarily have evidence of no relationship between GFOFAIL and MASTGOAL since power is unknown; rather the conclusion is that the relationship is unknown.

Fig. 7.1 Annotated output from R



```
> summary(M3)
Call:
mix(formula = MASTGOAL ~ 1 + (1 | CNTSCHID), data =
newdata,
   weights = c("W_FSTUWT", "W_SCHGRNRABWT"))

Variance terms:
 Level   Group       Name Variance Std. Error Std.Dev.
    2 CNTSCHID (Intercept)  0.07619    0.01632    0.276
    1 Residual            0.95453    0.02298    0.977
Groups:
 Level   Group n size mean wgt sum wgt
    2 CNTSCHID   142   148.7   21108
    1      Obs  4018    718.5 2886733

Fixed Effects:
          Estimate Std. Error t value
(Intercept)  0.22545   0.04571   4.932

lnl= -4049771.88
Intraclass Correlation= 0.07392
> summary(M4)
Call:
mix(formula = MASTGOAL ~ 1 + ESCS + Male + SCHSIZE_TH +
GFOFAIL +
   (1 | CNTSCHID), data = newdata, weights = c("W_FSTUWT",
   "W_SCHGRNRABWT"))

Variance terms:
 Level   Group       Name Variance Std. Error Std.Dev.
    2 CNTSCHID (Intercept)  0.08234    0.01737   0.2870
    1 Residual            0.91620    0.02212   0.9572
Groups:
 Level   Group n size mean wgt sum wgt
    2 CNTSCHID   142    148.7   21108
    1      Obs  4018    718.5 2886733

Fixed Effects:
            Estimate Std. Error t value
(Intercept)  0.6842304  0.0936302   7.308
ESCS         0.0880812  0.0206448   4.267
Male        -0.3641332  0.0347138 -10.490
SCHSIZE_TH   0.0920460  0.0461808   1.993
GFOFAIL     -0.0001962  0.0195705  -0.010

lnl= -3991608.83
Intraclass Correlation= 0.08246
>
```

Table 7.3 Summary of the results from M3 and M4

		Coef	SE
M3	Intercept	0.225	0.046
	School Var	0.076	
	Residual Var	0.955	
M4	Intercept	0.684	0.094
	ESCS	0.088	0.021
	MALE	−0.364	0.035
	SCHSIZE_TH	0.092	0.046
	GFOFAIL	0.000	0.020
	School Var	0.082	
	Residual Var	0.916	

Note Both models estimated with MASTGOAL outcome. Based on sample size of 4018 students nested within 142 schools

7.10 Conclusion

This tutorial was designed to assist applied researchers in analyzing large-scale assessment data, particularly PISA 2018 data, using multilevel modeling techniques. Aspects of complex survey design, including sampling weights and plausible values, were described and demonstrated using R statistical software. There are many variations on the multilevel model, such as multilevel structural equation models, multilevel logistic regression, and multilevel models for longitudinal data, that were not covered in the present demonstration. However, this tutorial should give the educational researcher a solid foundation on which to build applied analyses with PISA 2018 and other large-scale assessment data using multilevel modeling techniques.

Appendix: R syntax

```
#R syntax for demonstration analysis

###############################
#1. Import two packages for analysis of complex survey data and set working directory

#intsvy package will be used to import data & compute descriptive statistics
library(intsvy)

#WeMix package will be used for estimating multilevel models
#this package allows including of sampling weights in the model with the “mix” function
library(WeMix)

#set working directory (wd) where PISA data was saved
setwd("C:/Users/file path here")

###############################
#2. Save variable information and import data

#prints/saves variable labels and names of countries in a text file in working directory
pisa.var.label(folder=file.path(getwd(),"PISA 2018"),
        school.file="CY07_MSU_SCH_QQQ.sav",
        student.file="CY07_MSU_STU_QQQ.sav")

#selects & merges data; acheivement & weight variables selected by default
mydata <- pisa.select.merge(folder = file.path(getwd(), "PISA 2018"),
        school.file = "CY07_MSU_SCH_QQQ.sav",
        student.file = "CY07_MSU_STU_QQQ.sav",
        student = c("ESCS","ST004D01T","GFOFAIL","MASTGOAL"),
        school = c("W_SCHGRNRABWT","SCHSIZE"),
        countries = c("USA"))

###############################
```

```
#3. Data preparation

mydata$SCHSIZE_TH<-mydata$SCHSIZE/1000 #Size in thousands of students
mydata$Male<-mydata$ST004D01T #recode gender in direction of effect

##############################

#4. Descriptive statistics
#using functions from intsvy package
pisa.mean.pv(pvlabel="MATH",data=mydata)
pisa.mean(variable="MASTGOAL",data=mydata)
pisa.mean(variable="ESCS",data=mydata)
pisa.mean(variable="Male",data=mydata)
pisa.mean(variable="GFOFAIL",data=mydata)
pisa.mean(variable="SCHSIZE_TH",data=mydata)
pisa.table(variable="Male",data=mydata) #frequency distribution
pisa.ben.pv(pvlabel="MATH",data=mydata) #percents at proficiency levels
pisa.mean.pv(pvlabel="MATH",by="Male",data=mydata) #for plausible values
pisa.rho(variable=c("PV1MATH","MASTGOAL","ESCS","Male","GFOFAIL","SCHSIZE_TH"),
        data=mydata) #correlations
pisa.rho(variable=c("PV2MATH","MASTGOAL","ESCS","Male","GFOFAIL","SCHSIZE_TH"),
        data=mydata) #correlations
pisa.rho(variable=c("PV3MATH","MASTGOAL","ESCS","Male","GFOFAIL","SCHSIZE_TH"),
        data=mydata) #correlations

##############################
#5. Estimate multilevel models

#Math DV#
#listwise delete
newdata<-subset(mydata, select = c(PV1MATH,PV2MATH,PV3MATH,CNTSCHID,W_FSTUWT,W_SCHGRNRABWT,
        ESCS,Male,SCHSIZE_TH,GFOFAIL))
newdata <- na.omit(newdata)

#run models
```

```
M1a<-mix(PV1MATH~1+(1|CNTSCHID),data=newdata,
        weights=c("W_FSTUWT","W_SCHGRNRABWT"))
M1b<-mix(PV2MATH~1+(1|CNTSCHID),data=newdata,
        weights=c("W_FSTUWT","W_SCHGRNRABWT"))
M1c<-mix(PV3MATH~1+(1|CNTSCHID),data=newdata,
        weights=c("W_FSTUWT","W_SCHGRNRABWT"))
M2a<-mix(PV1MATH~1+ESCS+Male+SCHSIZE_TH+GFOFAIL+(1|CNTSCHID),data=newdata,
        weights=c("W_FSTUWT","W_SCHGRNRABWT"))
M2b<-mix(PV2MATH~1+ESCS+Male+SCHSIZE_TH+GFOFAIL+(1|CNTSCHID),data=newdata,
        weights=c("W_FSTUWT","W_SCHGRNRABWT"))
M2c<-mix(PV3MATH~1+ESCS+Male+SCHSIZE_TH+GFOFAIL+(1|CNTSCHID),data=newdata,
        weights=c("W_FSTUWT","W_SCHGRNRABWT"))

#MASTGOAL DV#
#listwise delete
newdata<-subset(mydata, select = c(MASTGOAL,CNTSCHID,W_FSTUWT,W_SCHGRNRABWT,
        ESCS,Male,SCHSIZE_TH,GFOFAIL))
newdata <- na.omit(newdata)

#run models
M3<-mix(MASTGOAL~1+(1|CNTSCHID),data=newdata,
        weights=c("W_FSTUWT","W_SCHGRNRABWT"))
M4<-mix(MASTGOAL~1+ESCS+Male+SCHSIZE_TH+GFOFAIL+(1|CNTSCHID),data=newdata,
        weights=c("W_FSTUWT","W_SCHGRNRABWT"))

##############################
#6. Combine results from multiple models to account for plausible values

#this demonstrates the procedure for the intercept from the empty model

#average of 3 intercept estimates represents combined coefficient
mean(M1a$coef,M1b$coef,M1c$coef)

#SE is computed with the following formula
#M is the number of PV; 3 for demonstration; PISA 2018 has 10 PV total
M<-3
mean(M1a$SE,M1b$SE,M1c$SE) + (1+(1/M))*var(c(M1a$coef,M1b$coef,M1c$coef))
```

References

Bailey, P., Kelley, C., Nguyen, T., & Huo, H. (2021). *WeMix: Weighted mixed-effects model using multilevel pseudo maximum likelihood estimation.* R package version 3.1.8. https://CRAN.R-project.org/package=WeMix

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Caro, D. H., & Biecek, P. (2017). intsvy: An R package for analyzing international large-scale assessment data. *Journal of Statistical Software, 81*(7), 1–44. https://doi.org/10.18637/jss.v081.i07

Ertem, H. Y. (2021). Examination of Turkey's PISA 2018 reading literacy scores within student-level and school-level variables. *Participatory Educational Research, 8*(1), 248–264. https://doi.org/10.17275/per.21.14.8.1

Ferrao, M. E., Costa, P. M., & Matos, D. A. S. (2017). The relevance of the school socioeconomic composition and school proportion of repeaters on grade repetition in Brazil: A multilevel logistic model of PISA 2012. *Large-Scale Assessments in Education, 5*(7), 1–13. https://doi.org/10.1186/s40536-017-0036-8

Lorah, J. A. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-scale Assessments in Education, 6*(8). https://doi.org/10.1186/s40536-018-0061-2.

Lorah, J. A. (2019). Estimating a multilevel model with complex survey data: Demonstration using TIMSS. *Journal of Modern Applied Statistical Methods, 18*(2), 2–14. https://doi.org/10.22237/jmasm/1604190360

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/methods/

Ma, X., Ma., L., & Bradley, K. D. (2008). Using multilevel modeling to investigate school effects. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 59–110). Information Age Publishing Inc.

OECD. (2009). *PISA data analysis manual: SPSS second edition.* Oecd.org.

OECD. (2021a, July 30). *PISA 2018 technical report, chapter 4: Sample design.* Oecd.org. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2021b, July 30). *How to prepare and analyse the PISA database.* Oecd.org. https://www.oecd.org/pisa/data/httpoecdorgpisadatabase-instructions.htm

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rogers, A. M., & Stoeckel, J. J. (2008). *NAEP 2008 arts: Music and visual arts restricted-use data files data companion* (NCES 2011–470). US Department of Education Institute of Education Sciences, National Center for Education Statistics.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publishing.

Dr. Julie Lorah is currently an assistant professor at Indiana University at Bloomington, Indiana, in the counseling and educational psychology department, where she teaches quantitative methods courses, including intermediate statistics, multivariate statistics, and multilevel modeling. She received her Ph.D. in educational psychology from the University of Washington. Her research interests involve the study and application of advanced statistical models, particularly the multilevel model, moderation model, and survival model, and methods for interpreting these models, including particularly measures of effect size. In addition, she is interested in issues of diversity within the field of statistics and statistics education.

Chapter 8
Multilevel Modelling of International Large-Scale Assessment Data

Anastasios Karakolidis, Vasiliki Pitsia, and Jude Cosgrove

Abstract It is indisputable that international large-scale assessments (ILSAs), such as the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA), play an important role in informing educational policies across countries. Such assessments provide rich but complex data. It is important to be aware of these complexities in order to analyse ILSA data correctly and interpret results appropriately. This chapter is an accessible introduction to the topic, providing a starting point for the application of multilevel modelling of ILSA data for research and policy. The chapter provides an introduction to key concepts and design features of ILSAs relevant to multilevel modelling (e.g., cluster sampling, weights, and plausible values) and considers issues from a practical perspective to support data preparation and the selection of modelling techniques and software.

Keywords Multilevel analysis · Large-scale assessments · PISA · TIMSS · PIRLS

8.1 Introduction

International large-scale assessments (ILSAs) are growing and evolving. For example, since the first cycle of the Programme for International Student Assessment (PISA) in 2000, the number of participating countries/jurisdictions has grown from 32 to 79 in PISA 2018, with 86 expected for PISA 2022. Alongside this expansion in scale, the complexity of the design of PISA and other ILSAs has increased, for example in moving from paper to online assessment, and with various enhancements to the scaling and analysis of data.[1]

[1] For example, compare OECD (2021), Chap. 9 to Adams and Wu (2002), Chap. 9.

A. Karakolidis (✉) · V. Pitsia · J. Cosgrove
Educational Research Centre, Dublin, Ireland
e-mail: anastasios.karakolidis@erc.ie

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_8

ILSAs have changed drastically since 1958, when early considerations of a study of "measured outcomes and their determinants within and between systems of education" were taking place (Husén & Postlethwaite, 1996, p. 129). That year is considered by some as the founding year of the International Association for the Evaluation of Educational Achievement (IEA) (Husén & Postlethwaite, 1996). Shortly afterwards, in 1961, the Organisation for Economic Co-operation and Development (OECD) was founded. The idea that international studies could provide information on "optimal conditions for human development that could be used as a basis for educational policy" (Kellaghan, 1996, pp. 143–144) has high intuitive, research and policy appeal. The organisers of the first international study commented: "If custom and law define what is educationally allowable within a nation, then educational systems beyond one's national boundaries suggest what is educationally possible." (Foshay et al., 1962, p. 7).

Various authors (e.g., Kellaghan & Greaney, 2001; Plomp et al., 2003) have described functions of ILSAs in terms of their potential relevance to the development of government policies on education:

1. Descriptive comparisons serving to identify aspects of a system that are at odds with others ('mirroring')
2. Benchmarking standards against which policymakers judge their education systems
3. Monitoring educational processes and outcomes over groups and time
4. Understanding differences between systems and groups to enable decisions about issues such as resource deployment and teaching and learning practices
5. Serving an "enlightenment function" by revealing assumptions about what schools or systems try to achieve through an analysis of what they actually achieve and a discussion about what is possible to achieve

Due, perhaps, to human tendencies to *compare* and *order*, media and policymakers tend, traditionally, to focus on descriptive comparisons and benchmarking standards arising from ILSAs (1 and 2 above). However, these do not provide the most useful policy and pedagogical information, particularly when we consider that most of the variation in achievement measures lies between individuals within countries, rather than between countries or jurisdictions. There are now multiple iterations of ILSAs. For example, PISA (since 2000, in its eighth cycle in 2022), the Trends in International Mathematics and Science Study (TIMSS, since 1995, eighth cycle in 2023), and the Progress in International Reading Literacy Study (PIRLS, since 2001, fifth cycle in 2021).[2] These data sources provide excellent opportunities to exploit the monitoring, understanding differences, and perhaps most importantly for policymakers, enlightenment functions of these studies (3, 4, and 5 above). Analyses of ILSA datasets within a multilevel modelling framework can inform these functions.

[2] Other similar studies are the International Civic and Citizenship Education Study (ICCS) and the International Computer and Information Literacy Study (ICILS).

Multilevel models offer a high level of flexibility in this regard since they may be used for both longitudinal and cross-sectional analysis, for both national and international analysis, and with various configurations of levels that take into account the clustered nature of the sampling design of ILSAs. Multilevel models that are well-conceived and sequenced can be used in the exploration of quite specific research questions and hypotheses. In this chapter, we show how various applications of the technique can provide important insights for monitoring cognitive and other outcomes over time, the utility of these applications in helping to understand differences between systems and groups, and their potential contribution to serving the enlightenment function of ILSAs.

Appropriate analysis of ILSA data is challenging, mainly due to issues emerging from the sophisticated design of these studies. Multilevel modelling helps us account for some of these issues, but it also has its own complexities. This chapter aims to provide a clear, non-technical, and practical explanation of the key design and measurement features of ILSAs, which, in turn, have implications for the manner in which multilevel models employing ILSA data are designed, analysed, and interpreted. Specifically, the chapter discusses key aspects of ILSAs—measures, number of levels, and weights—within a multilevel modelling framework. The chapter also includes practical considerations with respect to available software. The information provided in this chapter is directly relevant to a range of ILSAs such as PISA, TIMSS, PIRLS, ICCS, and ICILS, and is also relevant to the analysis of data from national assessments, which often have similar designs and procedures to those employed by the international studies (Greaney & Kellaghan, 2008).[3]

8.2 Sampling in ILSAs and the Use of Multilevel Models

ILSAs, such as PISA, TIMSS, and PIRLS, usually select their samples based on a two-stage process, involving schools as the primary sampling unit and either students or intact classes within the sampled schools as the secondary sampling unit (e.g., Martin et al., 2017, 2020; OECD, 2021). The clustered nature of these samples means that students within the same classes and/or schools are less likely to be independent of each other, as their knowledge, skills, and other attributes may be influenced by factors such as their classmates, teachers, school principals, and the overall school environment (Raudenbush & Bryk, 2002). Clustering constitutes a problem because many statistical models assume that cases are independent of each other. The degree of this clustering is commonly estimated using a statistic called the *intra-class correlation (ICC)*, which represents the proportion of the total variance in the outcome variable or a variable of interest that is attributable to the cluster(s) (Field, 2018). If a considerable proportion of the total variance in the outcome variable is

[3] For example, the National Assessments of Mathematics and English Reading (NAMER) in Ireland (Eivers et al., 2010) and the National Assessment of Educational Progress (NAEP) in the United States (National Center for Education Statistics, 2011).

attributable to the cluster(s) (i.e., if between-cluster variance is high), this should be taken into account in the analysis and, in particular, in the investigation of relationships between outcome and predictor variables stemming from different levels (e.g., student, class, school, district etc.) (Cohen et al., 2017). But even in cases of low ICC, it is recommended that the clustered nature of ILSA data is accounted for as there is still a hierarchical design effect that stems from the sampling design of these studies (Lai & Kwok, 2015; Snijders & Bosker, 2012). Ignoring the sampling design of these studies and, hence, violating the assumption of independence can result in an underestimation of the standard errors, which, in turn, may increase the risk of inflated Type I errors (i.e., mistaken rejection of the null hypothesis or false positive) (Field, 2018; OECD, 2009). Musca et al. (2011) used simulated data to show how Type I error rate varies according to the number of clusters, number of observations within clusters, and value of the ICC. They show that the risk of Type I error is highest when the number of higher-order units (clusters) is low, the number of observations within clusters is high, and the ICC is high. However, they recommend that, generally, clustered data should be analysed using multilevel modelling, commenting that "non-independence of data is not just a minor problem the researchers can afford to ignore. Quite to the contrary, the present simulations suggest that researchers will most likely draw incorrect conclusions if they fail to take the non-independence into account" (p. 4).

Multilevel modelling can take the clustering of the individuals into account, estimate the variation in the outcome variable that is attributable to differences within or between the clusters, and identify the factors at each level that are associated with this influence, while not underestimating the standard errors of the regression coefficients (Woltman et al., 2012). As Menezes et al. (2016) argued, multilevel modelling facilitates a more nuanced analysis of educational assessment data compared to other approaches, while considering the many potential levels of impact relevant to effective educational policy. Given that clustering of students in classes and/or schools is an inevitable reality in educational settings and in most ILSAs, it should be considered both in the relevant analyses and in policy-making processes.

8.3 Outcome Variables in ILSAs

8.3.1 Plausible Values

Besides their complex sampling designs, the psychometric consequences of ILSA test designs also need to be incorporated into the analysis process. Generally, due to time restrictions in ILSAs, each student is administered a subset of test items from the total item pool for each domain, with different groups of students answering different, although overlapping, sets of items. Consequently, individual student proficiencies are not fully observed and, therefore, the measurement of individual proficiency is achieved with a substantial amount of measurement error. Given this method of assessing students, and the fact that most ILSAs are designed to make population-level estimations, rather than accurately describe individual students' proficiencies based on their test scores, the imputation methodology of *plausible values* is often used (Rutkowski et al., 2010; von Davier et al., 2009). Plausible values constitute random draws from the distribution of scores that could be reasonably assigned to each individual (Wu, 2005).[4]

Since 2015, PISA moved from five to 10 plausible values for each of its scales, a transition that is likely to improve the accuracy of the estimations as more random draws are selected from the estimated ability distribution for each student (e.g., OECD, 2021). TIMSS and PIRLS have consistently used five plausible values to estimate student performance in mathematics, science, and reading and their subdomains across their administrations (e.g., Martin et al., 2017, 2020).

The fact that each assessed student is not assigned a single score but rather a set of plausible values has certain practical implications for the analysis of ILSA data in both single-level and multilevel contexts. The correct use of plausible values is sometimes overlooked in analyses of ILSA data, with the most common incorrect approaches being the use of one of or the mean of the plausible values as a single estimate of achievement. Such approaches generally underestimate standard errors of the generated statistics, which, in turn, is likely to lead to inflated Type I errors and incorrect conclusions.

In order to get unbiased estimates, researchers need to incorporate plausible values by conducting the analysis for each plausible value separately. The results of these analyses should then be averaged into a single set of point estimates and standard errors using formulas following Rubin's (1987) guidelines to account for imputation variability associated with generating plausible values. This process of combining the results of five or ten analyses (depending on the number of plausible values) adds further complexity to the analysis of ILSA data, but certain software programmes simplify the analysis of data with plausible values by automating the procedure. Additional information about software which can be used for this purpose can be found in the Software section of this chapter.

8.3.2 Continuous and Non-continuous Performance Outcomes

Performance data from ILSAs can be treated both as continuous (i.e., expressed in a numerical scale) and categorical (i.e., expressed as discrete performance categories).

[4] For further detail on the item response models and resultant plausible values used in ILSAs, readers are referred to the relevant technical documentation of the ILSA in question. The IEA and the OECD publish technical reports for each ILSA cycle; for example, see Chap. 12 of the PISA 2018 technical report (OECD, 2021), and Chaps. 11 and 12 of the TIMSS 2019 technical report (Martin et al., 2020).

For instance, using established levels of performance, the so-called proficiency levels in PISA and international benchmarks in TIMSS and PIRLS, ILSAs report student performance not only in continuous scores but also in categories of performance that indicate the performance level of students in each domain (see, for example, Mullis et al., 2020; OECD, 2021).[5] Treating students' performance as a continuous outcome is the most common way of analysing ILSA data, with many research studies using this approach within the context of single- or multilevel linear regression modelling. However, logistic models that treat student performance as a categorical outcome, with two or more categories, can also prove informative. The decision as to whether student performance is to be treated as a continuous or a categorical outcome should be informed by the purposes of a research study. For example, in a multilevel model having student socio-economic status as an explanatory variable and student science achievement as a continuous outcome variable, the focus of the analysis is on the achievement differences across socio-economic status scores. On the other hand, when achievement is being treated as a categorical variable (e.g., comparing low and non-low achievers), the results would focus on risk, in the sense that the model parameter estimates (expressed as odds ratios) will provide information on the extent to which students from lower socio-economic status are in greater risk of being low achievers.

The treatment of student performance as a categorical outcome is facilitated by the existing proficiency levels and international benchmarks in ILSAs, as is the prevention of independent researchers from arbitrarily separating levels of performance when analysing these data. However, multilevel logistic modelling, where the outcome is categorical, remains somewhat more complicated than linear models, where the outcome is continuous. The analysis of ILSA student performance as a categorical outcome using logistic regression modelling while taking all plausible values into account requires that certain steps are followed to prepare the data for analysis. These steps will vary depending on the particular ILSA being analysed. In the PISA databases, variables indicating the proficiency level to which students belong do not exist, but new variables can be created, using the cut-off points for each proficiency level of interest and for all plausible values. For example, to separate students into two groups, high achievers and non-high achievers, each student will have to be assigned the values 0 or 1 based on whether each plausible value estimate is below or above the established cut-off point for each respective domain. This should be done separately for each plausible value. Specifically, with 633 being the cut-off point indicating that students with a score of 633 or above are high achievers and those with scores 632.9 or below are non-high achievers, a student with plausible value estimates of 630, 631, 629, 634, and 633 would be assigned the values 0, 0,

[5] The way these levels are identified is through the use of specific cut-off points across the performance continuum for each plausible value in each domain. In each of the ILSAs, students scoring at certain levels in each domain, taking all plausible values into account, are identified as low, medium, or high achievers. Detailed descriptions of the skills that students are expected to demonstrate at each level of performance in each domain and ILSA, and further information about how the cut-off points for each level are set, can be found in the technical reports of ILSAs (see, for example, Martin et al., 2017, 2020; OECD, 2021).

0, 1, and 1; the student is considered a non-high achiever according to the first three plausible values, while according to the last two plausible values, they are considered a high achiever. This process is not required for TIMSS and PIRLS, as categorical variables corresponding to each plausible value exist in their databases. These binary variables can then be used as the outcome variables for different types of analyses that also take the plausible values into account.

It should also be noted that, along with their cognitive tests, ILSAs like PISA, TIMSS, and PIRLS, administer contextual background questionnaires to students, their parents, teachers, and school principals. While researchers may use variables measured as part of these questionnaires (e.g., students' confidence levels, family socio-economic status, bullying rates at school) as outcome variables in single-level or multilevel analyses, the most common outcome variable tends to be students' performance. To date, non-cognitive indices (measured through the contextual background questionnaires) have not been generated using plausible values, so the additional plausible values combination steps above would not apply to non-cognitive outcomes.

8.4 Configuration of Multilevel Models

In order to conduct meaningful and informative analyses, researchers should understand how the sampling design of the assessment study with which they are working may influence both the choice of the multilevel model configuration and the interpretation of the results. The most common configuration for multilevel analysis of ILSAs within one country is two-level, with students at level 1 and schools at level 2. While the sampling design of PISA fits this structure, the design of IEA studies, such as TIMSS and PIRLS, where intact classes, instead of individual students, are selected, allows for the introduction of another level that accounts for the between-class variance. In most cases, however, only one class per school is sampled and, therefore, there is no observed between-class variance within schools; in such cases, where only one class per school is sampled, it is recommended to implement two- rather than three-level models (Rutkowski et al., 2010).

It should be noted, though, that in TIMSS and PIRLS, countries have the option to select more than one class from each sampled school if they wish to increase their total student sample size or to provide more accurate estimates of the effects at the school level (Martin et al., 2020). Similarly, in national assessments, which often follow the design of international studies, countries may sample as many classes per school as they wish. When the number of sampled classes per school is not constant, with one class being sampled in some schools, and two or more classes being sampled in others, two-level models with students at the lower level and classes or schools at the upper level could be constructed. It should be noted, though, that in such models, the between-class and between-school variances are confounded as a proportion of the between-class variation will likely be due to differences between schools and vice versa. In such cases of non-constant sampling of classes, the application of three-level models (with students at level 1, classes at level 2, and schools at level 3) might also be sensible if the proportion of schools where more than one class has been sampled is high. When two or more classes per school have been sampled across all schools, the application of three-level models (with students at level 1, classes at level 2, and schools at level 3) is appropriate. In any case, with ILSAs like TIMSS and PIRLS, even when classes constitute a distinct level in multilevel analysis, it is advisable that teacher responses (coming from the teacher questionnaire) are treated and interpreted as student characteristics (e.g., Fishbein et al., 2021).[6] This is because, in these studies, teachers are not explicitly sampled and, therefore, inferences about teachers themselves are not appropriate (Fishbein et al., 2021; Rutkowski et al., 2010). However, this would not apply to studies where teachers are the main subject of the investigation and they are explicitly sampled to represent the broader population; an example of such a study is the Teaching and Learning International Survey (TALIS; OECD, 2019a).

It is not uncommon to see multilevel models, usually three-level ones, conducted across multiple countries, using country at the upper level; see, for example, OECD (2009, 2018). Such analyses can be useful for examining research problems with groups of countries that share common geographic or other characteristics (e.g., European Union or OECD countries). It could be also of interest to compare results across smaller groups of countries (e.g., Nordic countries) or regions within countries. However, a multilevel model with a country at the upper level may not be appropriate when there is an insufficient number of upper-level clustering units (i.e., countries); see Hox et al. (2018) and Kerkhoff and Nussbeck (2019) for recommendations on sample sizes in multilevel modelling. In any case, researchers should keep in mind that, countries, unlike schools, may not be representative of a broader population of countries and, therefore, the results of such analyses are relevant only to the cases included in the model.

If the aim of the analysis is to compare results across countries, then it is sensible to run separate, usually two-level, models within each country. After conducting the analysis separately for each country, a practical way of testing whether the model coefficients significantly differ across the examined countries is to enter the cases from all countries into a single model where the *country* is an explanatory dummy variable,[7] rather than a clustering factor, and test the statistical significance of the interactions between *country* and the other explanatory variables in the model; see, for example, Schütz et al. (2008) and van Daal et al. (2008).[8]

The results of analyses with multiple countries should always be interpreted with caution as, among others, the concept of school, structural features, such as streaming,

[6] It should be borne in mind that, in many cases, more than one teacher is linked to one class.

[7] The number of country dummies required in the model is $k - 1$, where k is the number of countries included in the analysis.

[8] This practice can be also applied to test how the relationships between explanatory and outcome variables change across different cycles of the same study within a country; see for example Karakolidis et al. (2021).

tracking, and selectivity, as well as other cultural, linguistic, and economic factors might significantly vary across different countries, which may, in turn, restrict any comparisons (Hanushek & Woessmann, 2005; OECD, 2009, 2013b). On the flip side, this natural variability and diversity, if adequately researched and appropriately incorporated into both the analyses and their resultant interpretations, can be exploited to inform the enlightenment function of ILSAs.

It should be noted that, even though multilevel models with students at level 1 are the most common ones, it is possible to run analyses that use schools at level 1 and countries at level 2. In any case, in order to select the optimal models for their multilevel analysis and appropriately interpret the results, researchers should carefully consider the aim of their research, understand the design of the assessment study with which they are working, and also consult the guidelines provided by the organisations responsible for each assessment study.

8.5 Sampling Weights in ILSAs

As explained earlier in this chapter, ILSAs involve complex designs and sampling strategies. As a result, each unit (i.e., student, class, or school) does not have the same chances of being selected to participate in a study. To control for this and generate results that can be generalised to the broader population, weights are used. The importance of using sampling weights to get robust estimates is referenced extensively in the literature (see Kim et al., 2013; Laukaityte & Wiberg, 2018; Mang et al., 2021; Rabe-Hesketh & Skrondal, 2006; Rutkowski et al., 2010), as ignoring weights can lead to biased results that may be considerably influenced by responses coming from certain groups of students.

The weights for each study reflect its sampling framework and design. For example, in PISA, where students are randomly selected within each one of the participating schools, two main sets of weights are computed and provided as variables in the databases: student weights and school weights. In PISA, the *total student weight* for student *i* in school *j* is computed as:

$$W_{ij} = \underbrace{wb_{ij} \times wa_{ij}}_{\text{final student weight}} \times \underbrace{wb_j \times wa_j}_{\text{final school weight}} \tag{8.1}$$

where wb_{ij} is the base weight for student *i* in school *j*, wa_{ij} is the non-response adjustment for student *i* in school *j*, wb_j is the base weight for school *j*, and wa_j is the non-response adjustment for school *j*. The combination of the base weights with the non-response adjustments gives the final student and school weights. The *total student weights* (W_FSTUWT) and the *final school weights* (W_SCHGRNRABWT) are included as variables in the PISA databases (OECD, 2021).[9]

[9] In PISA, the term *final weight,* rather than *total weight,* is used to refer to the student weights that incorporate the school weights (e.g., OECD, 2021). In this chapter, the terms *total* and *final* weights

Weighting becomes slightly more complex in studies like TIMSS and PIRLS, where intact classes, instead of individual students, are sampled and assessed. In such cases, the total student weights include the probability of selecting student *i* in class *j* and school *k* and the relevant non-response adjustments:

$$W_{ijk} = \underbrace{wb_{ijk} \times wa_{ijk}}_{\text{final student weight}} \times \underbrace{wb_{jk} \times wa_{jk}}_{\text{final class weight}} \times \underbrace{wb_k \times wa_k}_{\text{final school weight}} \tag{8.2}$$

where wb_{ijk}, wb_{jk}, and wb_k are the base weights and wa_{ijk}, wa_{jk}, and wa_k are the non-response adjustments for students (i), classes (j), and schools (k), respectively. The TIMSS and PIRLS databases include the *total student weights* (TOTWGT), the *overall* and *subject-specific teacher weights* (TCHWGT, MATWGT, and SCIWGT),[10] the *final school weights* (SCHWGT), and the sum of TOTWGT for all students within a school (STOTWGTU). Along with these weight variables, *house* (HOUWGT) and *senate* (SENWGT) weights, which are normalised for within- and across-country comparisons, are provided. TIMSS and PIRLS databases also include the base weights and the non-response adjustments for students, classes, and schools as individual variables, so that the final weights, as presented in formula 8.2, can be estimated at each level (i.e., student, class, school) (Fishbein et al., 2021). For example, the multiplication of the student base weighting factor (WGTFAC3) by the student non-response adjustment (WGTADJ3) gives the final student weight for TIMSS, as presented in formula 8.2 above. These computations for students, classes, and/or schools are necessary in the context of multilevel analyses of TIMSS and PIRLS data (Rutkowski et al., 2010).

The use of weights at the student level in single-level analyses is straightforward as the total student weights are appropriate. However, things get more complicated when it comes to multilevel modelling. Intuitively, it would appear correct to apply the pre-existing weights at each level of the analysis; for example, the total student weight at level 1 and the final school weight at level 2, in a two-level model with students at the first level and schools at the second level. However, this is not advisable because, as shown in the formulas above, the total student weights already incorporate the weights of the upper-level clusters (i.e., classes and/or schools) (Rabe-Hesketh & Skrondal, 2006). Therefore, for the purposes of multilevel analyses, it is important to be aware that some adjustments to existing weights may be required.

There are a number of different methods for computing the weights for multilevel analysis. Three of the most commonly used ones in the context of two-level models with ILSA data are the following:

Cluster weights: In this method, weights are scaled to add up to the cluster size. In the case of two-level models, with students at level 1 and schools at level 2, student

are used in line with the IEA studies; the former refers to the student weights that incorporate the school (and class) weights and the latter to the student weights that are free from the school (and class) weights (e.g., Martin et al., 2020).

[10] Teacher weights are not equivalent to class weights as the former are just total student weights divided by the number of teachers a student has (Rutkowski et al., 2010).

weights within each school are transformed to sum to the school sample size. At level 2, the final school weights, as those are provided in the ILSA databases, are used. Cluster weight for student *i* in school *j* is computed as:

$$W_{ij}^{*} = W_{ij} \frac{n_j}{\sum_j W_{ij}} \tag{8.3}$$

where W_{ij} is the *total* student weight, n_j is the student sample size in school *j*, and $\sum_j W_{ij}$ is the sum of total student weights in school *j* (Mang et al., 2021; Pfeffermann et al., 1998).

Clustersum weights: In this method, student weights (at level 1) add up to the number of sampled students within each school (similarly to the *Cluster* method). However, at the school level (level 2), instead of the final school weights, the sum of total student weights within each school is computed and used (Mang et al., 2021). This weight scaling approach has been used in multilevel models in the OECD reports since PISA 2012 (see OECD, 2013a, 2016, 2019b).

Withincluster weights: In this method, student weights (at level 1) are computed to be independent from school weights, while school weights, as those are provided in the ILSA databases, are used in level 2 (Mang et al., 2021).

To extract the final student weights, as those are presented in formula 8.1, when analysing PISA data, the total student weights are divided by the final school weights (see Table 8.1). Even though this approach can be used when analysing TIMSS and PIRLS data, these IEA databases include base weights that can be used for the computation of the final weights at each level, as shown in formula 8.2. For three-level models with students at level 1, classes at level 2, and schools at level 3, the final weights at each level can be used. However, when two-level models are applied, the final class weights should be combined with the final student weights (in models with schools at the upper level; see example in Table 8.1) or with the final school weights (in models with classes at the upper level).

Table 8.1 summarises the three weighting methods for two-level models. For various permutations of three-level models, the logic is the same as for two-level models.

A review of the relevant literature shows that there is a tendency for the *Cluster* and *Clustersum* methods to be used in multilevel analyses of PISA data (OECD, 2019b; Sempé, 2021), while when IEA data (e.g., TIMSS and PIRLS) are used, the *Withincluster* method tends to be more popular (e.g., Ersan & Rodriguez, 2020); however, this does not mean that the *Cluster* and *Clustersum* methods cannot be used for IEA data, or that the *Withincluster* method cannot be used for PISA data. In fact, Mang et al. (2021) examined the robustness of these and other weighting approaches using the PISA 2015 data for Germany.[11] The authors conducted multiple simulations to compare model estimates using different weight scaling methods across different variance distribution scenarios: original between-school variance ($ICC = 0.52$), low

[11] The nine weighting approaches Mang et al. (2021) compared in their study were: (i) no weights, (ii) unscaled weights, (iii) only student weights, (iv) only school weights, (v) house weights, (vi) cluster weights, (vii) ecluster weights, (viii) clustersum weights, and (ix) withincluster weights.

Table 8.1 Summary of methods for applying weights to two-level models (level 1: students, level 2: schools)

Method	Description	Computation—PISA	Computation—TIMSS & PIRLS
Cluster	*Level 1:* student weights are scaled to add up to the cluster size *Level 2:* the readily available final school weights are used	*Level 1:* formula 8.3 *Level 2:* use *W_SCHGRNRABWT* variable	*Level 1:* formula 8.3 *Level 2:* use *SCHWGT* variable
Clustersum	*Level 1:* student weights are scaled to add up to the cluster size *Level 2:* the sum of the total student weights within each school is used	*Level 1:* formula 8.3 *Level 2:* compute the sum of total student weights within each school	*Level 1:* formula 8.3 *Level 2:* use *STOTWGTU* variable
Withincluster	*Level 1:* the student weights, independent from school weights, are used *Level 2:* the readily available final school weights are used	*Level 1:* divide *W_FSTUWT* by *W_SCHGRNRABWT* *Level 2:* use *W_SCHGRNRABWT* variable	*Level 1:* combine the student and the class weights via multiplication ($WGTFAC3 \times WGTADJ3 \times WGTFAC2 \times WGTADJ2$) or divide *TOTWGT* by *SCHWGT* *Level 2:* use *SCHWGT* variable

between-school variance ($ICC = 0.05$), and high between-school variance ($ICC = 0.79$). Mang et al. (2021) concluded that the use of weights, especially at the school level, is needed to get unbiased estimates. Specifically, weighting approaches *Cluster* and *Only school weights* (with no weights at the student level) were found to be among the least biased ones across all models and scenarios. In light of these findings, the authors argued that the use of weights only at the school level (as those are provided in the ILSA databases) might suffice, and they recommended this weighting approach due to its simplicity.

In any case, it would be worth comparing different weighting approaches in initial exploratory analyses to identify the most informative one; the use of informative weights usually leads to less biased estimates, something that is evident by the increased standard errors (Kim et al., 2013). For instance, someone could evaluate whether the use of student weights, on top of the school weights, is necessary in a two-level model by comparing the results of analysis with and without the use of weights at level 1. If the use of weights at the student level leads to increased standard errors of the estimates in the model, this indicates that student weights are informative and should be used. Some analysis software programmes can facilitate these comparisons by automatically scaling weights at each level so that the researchers do not have to do it manually; for example, *Cluster* is the default method for scaling weights at level 1 in *Mplus* (Muthén & Muthén, 2017).

For researchers interested in analyses using countries as the upper level, it should be noted that it is often desirable that each country contributes equally to the analysis so that the results are not dominated by countries with larger sample sizes and/or populations. In such cases, weights within each country should be normalised so that their sum remains the same for all countries included in the analysis. As mentioned above, TIMSS and PIRLS provide a normalised version of the total student weights (senate weights), the sum of which is constant across countries. These weights are suitable for two-level models with students at level 1 and countries at level 2; however, when researchers wish to run a three-level model with schools at level 2 to also account for the between-school variance within each country, then weights at the student and school level should first be scaled according to one of the approaches described in this chapter, and subsequently normalised so that their sum is constant across countries. In practice, this can be achieved if the sum of the weights at each level per country is constant and equal to the total number of cases divided by the number of countries (OECD, 2009). The following formula shows how the student weights in a three-level model can be converted so that their sum is constant across countries:

$$W_{ijk}^{**} = \frac{W_{ijk}^{*}}{N_k} \times \frac{n}{c} \tag{8.4}$$

where W_{ijk}^{*} is the normalised or scaled weight for student i in school j of country k, N_k is the number of cases (i.e., students) in the population of country k, n is the total number of sampled students across all countries included in the analysis, and c is the number of countries included in the analysis. This formula can be adjusted to also convert school or class weights.

As a final note, it should be highlighted that weights, as provided in ILSA databases, often sum to the size of the populations. This can lead to underestimation of standard errors and, hence, to biased results (increased rates of Type I errors). Therefore, independently of whether single-level or multilevel analysis is conducted, it is recommended that the weights are always normalised (i.e., divided by the mean of the weights), so that they sum to the sample size, rather than to the population.

Some software programmes handle this issue by default (e.g., *Mplus*), while others (e.g., *SPSS*) do not and, therefore, normalised weights should be computed.

8.6 Software

As mentioned earlier in this chapter, analysing ILSA data using multilevel modelling is a complex process due to the particularities of the data and the complexity of the analysis itself. However, in recent years, it has become more accessible to researchers due to the availability of relevant resources and specialised software. Although a detailed overview of all the software that can be used to analyse ILSA data in a multilevel context is beyond the scope of this chapter, some of the software that are most frequently used by the organisations responsible for the major ILSAs and by independent users of ILSA data are listed.

Software developed specifically for multilevel modelling, such as *HLM* (Garson, 2013) and *MLwiN* (Rasbash et al., 2020), as well as other more general-purpose software, such as *Mplus* (Muthén & Muthén, 2017), *R* (R Core Team, 2020), *Stata* (StataCorp, 2021), and *SAS* (SAS Institute Inc., 2018), can be used for multilevel modelling of ILSA data. *Mplus* and *HLM* appear to be particularly popular among researchers analysing ILSA data in a multilevel context as they support a range of model specifications (e.g., linear and logistic multilevel modelling), while they can also handle weights at different levels of the analysis and plausible values. Open-source software programmes, such as *R*, that allow users to use, amend, and distribute the source code and are partially or fully available for free are also becoming increasingly popular. Specifically, more packages designed for analysing ILSA data in *R* are becoming available with some of them allowing for multilevel analysis; see, for example, *BIFIEsurvey* (BIFIE et al., 2019), *RALSA* (Mirazchiyski & INERI, 2021), and *EdSurvey* (Bailey et al., 2021). The *IEA IDB Analyzer* (IEA, 2021) is another software that has been specifically designed for the analysis of ILSA data, and although it does not support multilevel modelling per se, it can adjust for sampling error via the use of the replicate weights.

All these software and packages, accompanied by relevant literature, discussion forums, product support(s), and professional development courses, have helped researchers to overcome the challenges of employing multilevel modelling for ILSA data and conduct advanced analyses that take into account the complex nature of these data. It must be noted here that any section on software may quickly become out of date; hence, the information presented here is based on the current state of the art and the authors' knowledge. Researchers interested in conducting such analyses with ILSA data should always consult the latest documentation on the software they use and ensure that their analyses adhere to certain guidelines outlined in the technical reports of each assessment (e.g., Martin et al., 2020; OECD, 2021), the data analysis manuals or user guides for the databases of each of the assessments (e.g., Fishbein et al., 2021; OECD, 2009), and other relevant documentation (e.g., Rutkowski et al., 2010; von Davier et al., 2009).

8.7 Summary

This chapter provided an introduction to the key features of ILSAs which have implications for their analyses within a multilevel modelling framework. ILSAs are now significantly—almost unrecognisably—developed from their inception in the early 1960s. Although descriptive comparisons and benchmarking standards have been the most common uses of ILSA data by policymakers and reported on by the media, multilevel modelling provides an appropriately sophisticated tool that enables more insightful uses of these datasets, such as monitoring of educational standards over groups and over time, understanding of differences between systems and groups to inform policy and practice, and exploitation of the enlightenment function of ILSAs, especially when used in conjunction with comparative and national policy analysis.

Regarding the technical features of ILSAs, we have shown how their sampling design should be considered by researchers to guide decisions about the configuration of the levels within the model, and also how the natural diversity of countries and systems acts as both an opportunity for insight, as well as a risk, if not considered carefully in both the multilevel analysis and its interpretation. We also provided some initial guidance on and consideration of the consequences of how ILSAs' test designs and plausible values can be correctly handled in the analysis. We drew attention to the fact that cognitive test outcomes can be treated as either continuous or categorical, and that the choice of test outcome measures should be matched to the research questions and overall aims of the analyses and their implications for research, policy, and/or practice. Finally, we described ways in which sampling weights may be optimally incorporated into multilevel models. Overall, we encourage researchers to conduct exploratory analyses, such as comparisons of different configurations of weights, prior to "settling" on the specifics of the model.

Fortunately, a range of software tools is available for the appropriate and efficient analysis of ILSAs and, in this chapter, we briefly described the features of the main software tools available, noting that the most recent versions of software and accompanying documentation should be used when selecting the optimal software tools for analysis.

The information provided in this chapter is directly applicable to well-established ILSAs, such as PISA, TIMSS, and PIRLS, but it can also be very relevant to other large-scale assessments, such as national studies, that often follow similar complex designs. Researchers should always consider the aim of their research, the design of the assessment study they are working with, as well as the particularities of the analysis software they use to configure appropriate and informative multilevel models. In all cases, though, the *what* and the *why* of the analysis in terms of its contributions to research and policy should precede the more technical *how*.

Acknowledgements The authors are indebted to Alice Duggan for proofreading the chapter.

References

Adams, R., & Wu, M. (Eds.) (2002). *PISA 2000 technical report*. PISA, OECD Publishing. https://doi.org/10.1787/9789264199521-en

Bailey, P., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., Nguyen, T., Xie, Q., Yu, J., Zhang, T., Buehler, E., & Lee, S. (2021). *EdSurvey: Analysis of NCES education survey and assessment data (R package version 2.7.0)*. https://cran.r-project.org/package=EdSurvey

BIFIE, Robitzsch, A., & Oberwimmer, K. (2019). *BIFIEsurvey: Tools for survey statistics in educational assessment (R package version 3.3–12)*. https://cran.r-project.org/package=BIFIEsurvey

Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education* (8th ed.). Routledge.

Eivers, E., Clerkin, A., Millar, D., & Close, S. (2010). *The 2009 National Assessments technical report*. Educational Research Centre.

Ersan, O., & Rodriguez, M. C. (2020). Socioeconomic status and beyond: A multilevel analysis of TIMSS mathematics achievement given student and school context in Turkey. *Large-Scale Assessments in Education*, *8*(15). https://doi.org/10.1186/s40536-020-00093-y

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database.* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries.* UNESCO Institute for Education.

Garson, G. D. (2013). Introductory guide to HLM with HLM 7 software. In G. D. Garson (Ed.), *Hierarchical linear modeling: Guide and applications* (pp. 55–96). SAGE Publications Inc. https://doi.org/10.4135/9781483384450.n3

Greaney, V., & Kellaghan, T. (2008). *Assessing national achievement levels in education.* World Bank. https://hdl.handle.net/10986/6904

Hanushek, E. A., & Woessmann, L. (2005). *Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries* (IZA DP No. 1901).

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Husén, T., & Postlethwaite, T. N. (1996). A brief history of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education: Principles, Policy & Practice, 3*(2), 129–141. https://doi.org/10.1080/0969594960030202

IEA. (2021). *Help manual for the IEA IDB analyzer (Version 4.0).* https://www.iea.nl

Karakolidis, A., Duggan, A., Shiel, G., & Kiniry, J. (2021). Examining educational inequalities: Insights in the context of improved mathematics performance on national and international assessments at primary level in Ireland. *Large-Scale Assessments in Education*, *9*(5). https://doi.org/10.1186/s40536-021-00098-1

Kellaghan, T. (1996). IEA studies and educational policy. *Assessment in Education: Principles, Policy & Practice, 3*(2), 143–160. https://doi.org/10.1080/0969594960030203

Kellaghan, T., & Greaney, V. (2001). The globalisation of assessment in the 20th century. *Assessment in Education: Principles, Policy & Practice, 8*(1), 87–102. https://doi.org/10.1080/09695940120033270

Kerkhoff, D., & Nussbeck, F. W. (2019). The influence of sample size on parameter estimates in three-level random-effects models. *Frontiers in Psychology, 10.* https://doi.org/10.3389/fpsyg.2019.01067

Kim, J. S., Anderson, C. J., & Keller, B. (2013). Multilevel analysis of assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-Scale assessment: Background, technical issues, and methods of data analysis.* Chapman and Hall/CRC Press. https://doi.org/10.1201/b16061

Lai, M. H. C., & Kwok, O. (2015). Examining the rule of thumb of not using multilevel modeling: The "design effect smaller than two" rule. *The Journal of Experimental Education, 83*(3), 423–438. https://doi.org/10.1080/00220973.2014.907229

Laukaityte, I., & Wiberg, M. (2018). Importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communications in Statistics—Theory and Methods, 47*(20), 4991–5012. https://doi.org/10.1080/03610926.2017.1383429

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education*, *9*(6). https://doi.org/10.1186/s40536-021-00099-0

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Menezes, I. G., Duran, V. R., Mendonça Filho, E. J., Veloso, T. J., Sarmento, S. M. S., Paget, C. L., & Ruggeri, K. (2016). Policy implications of achievement testing using multilevel models: The case of Brazilian elementary schools. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01727

Mirazchiyski, P., & INERI. (2021). *RALSA: R analyzer for large-scale assessments (R package version 1.0.2)*. https://cran.r-project.org/package=RALSA

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science.* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: Impact of intraclass correlation and sample size on Type-I error. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00074

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

National Center for Education Statistics. (2011). *Overview of the NAEP assessment design.* NAEP Technical Documentation. https://nces.ed.gov/nationsreportcard/tdw/overview/

OECD. (2009). *PISA data analysis manual: SPSS second edition.* PISA, OECD Publishing. https://doi.org/10.1787/9789264056275-en

OECD. (2013a). *PISA 2012 results: Excellence through equity (Volume II): Giving every student the chance to succeed*. PISA, OECD Publishing. https://doi.org/10.1787/9789264201132-en

OECD. (2013b). *PISA 2012 results: What makes schools successful (Volume IV): Resources, policies and practices.* PISA, OECD Publishing. https://doi.org/10.1787/9789264201156-en

OECD. (2016). *PISA 2015 results (Volume II): Policies and practices for successful schools.* PISA, OECD Publishing. https://doi.org/10.1787/9789264267510-en

OECD. (2018). *Effective teacher policies: Insights from PISA.* PISA, OECD Publishing. https://doi.org/10.1787/9789264301603-en

OECD. (2019a). *TALIS 2018 technical report*. OECD Publishing.

OECD. (2019b). *PISA 2018 results (Volume III): What school life means for students' lives.* PISA, OECD Publishing. https://doi.org/10.1787/acd78851-en

OECD. (2021). *PISA 2018 technical report*. PISA, OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *60*(1), 23–40.

Plomp, T., Howie, S., & McGaw, B. (2003). International studies of educational achievement. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation. Kluwer*

International Handbooks of Education (Vol. 9, pp. 951–978). Springer. https://doi.org/10.1007/978-94-010-0309-4_53

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, 169*(4), 805–827. https://doi.org/10.1111/j.1467-985X.2006.00426.x

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2020). *A user's guide to MLwiN, v3.05*. Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). SAGE Publications, Inc.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons Inc. https://doi.org/10.1002/9780470316696

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

SAS Institute Inc. (2018). *SAS/STAT® 15.1 user's guide*. SAS Institute Inc.

Schütz, G., Ursprung, H. W., & Wößmann, L. (2008). Education policy and equality of opportunity. *Kyklos, 61*(2), 279–308. https://doi.org/10.1111/j.1467-6435.2008.00402.x

Sempé, L. (2021). School-level inequality measurement based categorical data: A novel approach applied to PISA. *Large-Scale Assessments in Education*, *9*(9). https://doi.org/10.1186/s40536-021-00103-7

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE.

StataCorp. (2021). *Stata base reference manual: Release 17*. Stata Press.

van Daal, V., Begnum, A. C., Solheim, R. G., & Adèr, H. (2008). Nordic comparisons in PIRLS 2006. *3rd IEA International Research Conference (IRC-2008).*

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 9–36.

Woltman, H., Feldstain, A., Mackay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52–69. https://doi.org/10.20982/tqmp.08.1.p052

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Dr. Anastasios Karakolidis is a Research Associate at the Educational Research Centre, Ireland, and mainly works on the National Assessments of Mathematics and English Reading. He has been involved in various national and international research projects, has given lectures on research methods, statistical analysis, assessment, and measurement, and has published his research in peer-reviewed academic journals. He has extensive experience in analyzing and reporting large-scale assessment data, and he is also particularly interested in validity, measurement, and technology-based assessments.

Dr. Vasiliki Pitsia is a Research Assistant at the Educational Research Centre, Ireland. She holds a BEd from the University of Ioannina, Greece, an MSc in Quantitative Methods and Statistical Analysis in Education from Queen's University Belfast, United Kingdom, and a PhD in Assessment from Dublin City University, Ireland. Vasiliki has worked as a researcher, data analyst, and psychometrician on various projects in Ireland and Greece and as a statistical consultant at the World Bank Group. She has also given lectures on research methodology, measurement, assessment, and statistics to postgraduate students and staff in academic institutions in Ireland, Greece, and Cyprus. Her research interests include research methodology, statistical analysis, psychometrics, measurement, and assessment.

Dr. Jude Cosgrove Cosgrove is the Chief Executive Officer of the Educational Research Centre (ERC, Dublin, Ireland) since 2018. She has 25 years of experience in development, oversight, and reporting on large-scale national and international assessments and has led on or contributed to various national evaluations, having taken multiple roles including PISA National Project Manager and PISA Governing Board representative. She is currently the Irish representative on the PIAAC Board of Participating Countries and sits on the IEA's General Assembly. She has a keen interest in digital technologies in education and making better use of educational research data and studies to inform policy.

Chapter 9
Transparency and Replicability of Multilevel Modeling Applications: A Guideline for Improved Reporting Practices

Wen Luo, Eunkyeng Baek, and Haoran Li

Abstract Multilevel modeling (MLM) is a statistical technique for analyzing clustered data. Given the complexity of multilevel models, it is crucial for researchers to provide complete and transparent descriptions of the data, statistical analyses, and results. A recent systematic review of the reporting practices in MLM applications in education and psychology showed that the reporting practices still lack clarity and completeness in some areas, including reliability and validity of multilevel measures, model specifications, description of missing data mechanisms, power analyses, assumption checking, model comparisons, and effect sizes (Luo et al, 2021). In this chapter, we aim to provide a guideline for improved reporting practices in the identified areas to enhance the transparency and replicability of MLM applications. We will offer suggestions for what and how to report MLM results in those areas, use examples from real life research to illustrate the principles and guidelines, and provide readers with a checklist to describe the main points that should be thoroughly checked and clearly conveyed in reports when applying MLM.

Keywords Multilevel modelling · Multilevel research · Reporting practice

9.1 Introduction

Multilevel modeling (MLM) is a statistical technique for analyzing clustered or nested data, such as students nested within schools, or repeated measures nested within individuals in longitudinal studies. With the rapid development of MLM techniques and the accompanying computer programs, there has been a significant increase in the quantity and complexity of MLM applications. For example, MLM has been applied to various research designs (e.g., cluster randomized controlled trials, single-case experimental designs, non-experimental designs, or meta-analyses), different types of outcomes (e.g., continuous or categorical), and various multilevel data structures (e.g., strictly nested, cross-classified, or multiple-membership).

W. Luo (✉) · E. Baek · H. Li
Texas A&M University, College Station, Texas, USA
e-mail: wluo@tamu.edu

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_9

Given the complexity of multilevel models, it is crucial for researchers to provide complete and transparent descriptions of the data, statistical analyses, and results, so that “scientific claims can be clearly understood, assessed, and evaluated by the reader” and “the work can be replicated with reasonable accuracy” (Appelbaum et al., 2018, p. 23).

Although the initial guidelines for reporting MLM applications were proposed more than 10 years ago (Dedrick et al., 2009), a recent systematic review of MLM applications based on 301 articles from 19 journals in education and psychology showed that the majority of the studies still lacked clarity and completeness in reporting the psychometrics (i.e., reliability and validity) of multilevel measures, model specifications, description of missing data mechanisms, power analyses, assumption checking, model comparisons, and effect sizes (Luo et al., 2021).

Given the gaps between what was recommended in the methodological literature and what was reported in the applied multilevel research, we believe a guideline for improved reporting practices, especially in the areas where poor practices persist, would be beneficial to applied researchers. Using the most recent reporting standards for quantitative research by the American Psychological Association (APA; Appelbaum et al., 2018) as a general framework, we offer suggestions for *what* and *how* to report MLM research in the sections of hypotheses, methods, results, and discussion. We use examples motivated by real world studies to illustrate the principles and guidelines. Finally, we provide readers with a checklist to describe the main points that should be thoroughly checked and clearly conveyed in reporting MLM applications.

The examples are based on the Early Childhood Longitudinal Studies—Kindergarten Class of 2011 (ECLS-K: 2011), a longitudinal study that followed a nationally representative cohort of kindergarteners in 2011 through the fifth grade. The ECLS-K data has a multilevel structure and a rich collection of variables measured at various levels (e.g., student, parent, teacher, and schools). It has been widely used to study the associations between a wide range of family, school, community, and individual factors and students’ academic and social emotional development. In Example 1, the main purpose is to examine the association between students’ achievement and parental involvement using cross-sectional data in which students are nested within schools. In Example 2, the main purpose is to examine the development of self-control using longitudinal data in which repeated measures are nested within students and students nested within schools.

The remaining chapter is organized in the order of sections appearing in a typical quantitative research paper.

9.2 Statement of Research Questions and Hypotheses

As an initial step of the reporting process, researchers should describe the purpose of a study and clearly state the research questions. A clear statement of the research question is a fundamental step because it will lead to the development of an appropriate analytical strategy (e.g., choice of study designs, identification of appropriate measures, and specification of statistical models) that is aligned with the research question.

There are various types of research questions that can be investigated using MLM. For example, in Example 1 where a cross-sectional design is used, the main purpose is to examine the association between students' achievement and parental involvement. In this example, researchers can set a primary research question as "What is the *average* relationship between students' reading achievement and parental involvement across all schools?" In addition to the average relationship, researchers may also be interested in examining if the relationship between students' reading achievement and parental involvement varies across schools as a secondary question. In this case, a research question can be stated as "To what degree does the relationship between the students' reading achievement and parental involvement vary across schools?" Researchers may want to explore potential student factors (e.g., gender, social economic status) and school factors (e.g., school type, school size) that may change the relationship between students' reading achievement and parental involvement. For example, an exploratory question can be stated as "Does the relationship between student achievement and parent involvement vary depending on student social economic status?" Such questions involve the test of specific moderation (or interaction) effects that need to be explicitly stated. Furthermore, MLMs allow researchers to partition the effect of parental involvement into the between-school and the within-school effects. For the between-school effect, parental involvement is conceptualized as a school's organizational context, representing the patterns of the parent–school relationship. For the within-school effect, parental involvement is conceptualized as an individual student's parent involvement relative to the average level of parental involvement in the school. Specific research questions need to be stated for the between-school and within-school effect if they are of interest.

In Example 2 where a longitudinal design is used, the main purpose is to examine the development of self-control over time. Potential research questions may include "How does students' self-control ability develop over the elementary years from kindergarten to fifth grade?", "What is the initial status of the self-control ability on average (i.e., self-control ability in kindergarten) and how much does it vary across students?", "What is the growth rate of the self-control ability on average and how much does it vary across students?", and "What individual and/or school factors may be associated with the variations in individual student's growth trajectories?"

As illustrated above, MLM is flexible to answer various research questions afforded by different types of design. Below is a non-exhaustive list of the questions that can be addressed by MLM with more generality. For cross-sectional designs, MLM can address questions including, but not limited to, (1) the mean and variance of an outcome within and across clusters, (2) average relationship between predictors and an outcome across clusters, (3) variation in the relationship between predictors and an outcome across clusters, (4) moderating effects of level-1 or higher-level predictors on the relationship between predictors and an outcome. For the longitudinal design, MLM can further address questions related to (1) the average initial status and the rate of change (e.g., growth) on an outcome, (2) the form of the change (e.g., linear, nonlinear), (3) variation in the initial status and the rate of change, (4) relationship between initial status and the rate of change, (5) potential impact of level-1 or higher-level factors on the initial status and the rate of change. Given the many questions that can be addressed in a study, it is recommended that researchers explicitly state all of the research questions examined in a study, differentiating primary vs. secondary research questions, a priori vs. post hoc questions, and confirmatory vs. exploratory questions.

9.3 Description of the Sampling Procedures

When researchers describe collected data for the study, it is important to report whether the sample is a probability or a nonprobability sample (e.g., convenient sample) and if a probability sample is obtained, what sampling methods are employed to achieve the probability sample. For example, the ECLS-K study employed a multi-stage sampling design (i.e., selection of counties, selection of schools within selected counties, and selection of students within selected schools) to obtain a nationally representative probability sample. At each stage, various types of sampling methods, such as oversampling, clustered, and stratified methods, were also employed to make a representative probability sample (Najarian et al., 2019).

It is possible that each sampling stage may employ a different sampling procedure. For example, researchers may collect a nonprobability sample of school level in which only schools that volunteered to participate are included in the study, and then collect a probability sample of students within participating schools. In this case, researchers should report the sampling methods applied at each level separately.

Multilevel data collected using a complex sampling design often yields sampling units with different probabilities of being selected (Thomas & Heck, 2001). For example, in the ECLS-K sampling design, Asian, Native Hawaiian, and other Pacific Islander children were oversampled which made those children with a higher chance of being selected than children of other races. In this case, sampling weights are often used to obtain accurate estimates of population parameters (Asparouhov, 2006; Pfeffermann, 1993; Pfeffermann et al., 1998). In the ECLS-K data, various sampling weights have been applied for school-level (e.g., school administrator questionnaire) and child-level (e.g., child assessment, teacher-level questionnaire, child-level teacher questionnaire, parent interview, etc.) variables to compensate for differential probabilities of selection at each sampling stage as well as to adjust for the effect nonresponse.

Researchers should explicitly report whether sampling weights are employed or not, and if multiple sets of sampling weights are used, researchers should be clear about which sampling weights are applied to which level or which variables in each level. However, according to Luo et al. (2021)'s study, among the studies that used probability sampling, more than half of them did not explicitly report whether sampling weights were used or at which level sampling weights were applied. Hence, it is important to provide sufficient information regarding the sampling procedure and sampling weights used for the analysis in order to inform the limitation of the sample and to allow other researchers to replicate the study.

9.4 Sample Size, Power, and Precision

In multilevel data, sample sizes vary depending on the level of units. It is an essential practice for researchers to report the total sample size at each level as well as the variation of cluster size across clusters. For example, in Example 1 where students (level-1) are nested within schools (level-2), the total number of students and the number of schools should be reported separately. In addition, the average number of students per school (i.e., cluster size) and the variation of cluster size (e.g., minimum and maximum cluster size, the interquartile range of cluster size, or standard deviation of cluster size) should be reported. In Example 2 where observations (level-1) are nested within students (level-2) and students within schools (level-3), the number of observations at each time point, the number of students, the number of schools, and the number of observations per student are all essential information to be reported. In addition, if there is attrition at any time point, researchers should report the attrition rate at each wave. More details on reporting attrition are provided in the section of Missing Data Treatment.

In multilevel research, reporting the sample size at each level is particularly important because it is related to important statistical aspects, such as a model specification, statistical power, and the precision of the parameter estimates. For example, although the average relationship between student's reading achievement and parental involvement can be accurately estimated in a sample with a sufficiently large number of students, the variance of the slope of parent involvement across schools may not be accurately estimated and the statistical test for any school-level predictor may lack power or statistical precision if there are too few schools (Example 1).

In addition to reporting sample sizes, researchers should also provide an assessment as to whether the obtained sample size is adequate for estimating and testing the effects of interest. There are some conventional guidelines for adequate sample sizes in MLM analyses. For example, in a two-level model, at least 30 level-2 units with 30 level-1 units per level-2 unit are recommended to accurately estimate fixed effects of level 1 or level 2 predictors (30/30 rule) (Kreft, 1996). For cross-level interaction effects, the 50/20 rule is recommended. The 100/10 rule is suggested for variance component estimates (Hox, 1998). More recent studies indicated that certain correction methods, such as Kenward- Roger adjustment, and Bayesian estimation are useful for small sample sizes to obtain accurate estimates and statistical inferences (Baek et al., 2020; Baek & Ferron, 2020; Ferron et al., 2009; McNeish, 2016; McNeish, 2017).

It should be noted that even though the sample size meets the minimum requirement of the conventional guidelines, a study may still lack power when testing certain effects. Thus, it is recommended that researchers conduct power analyses to determine the required sample size to achieve an adequate level of power rather than relying on the general sample size guideline (Dedrick et al., 2009; Ferron et al., 2008; Raudenbush, 1997; Raudenbush & Liu, 2000). Several approaches, such as formula-based and simulation-based approaches (Crainiceanu & Ruppert, 2004; Gastanaga et al., 2006; Mathieu et al., 2012), as well as specialized software, such as Optimal Design (Raudenbush et al., 2011), PINT (Snijders et al., 1996), PowerUp! (Dong & Maynard, 2013), and the R package *longpower* (Donohue et al., 2013) and *simr* (Green & MacLeod, 2016) are available to conduct multilevel power analyses.

9.5 Psychometrics

When a measurement instrument (e.g., a survey or a test) is used to obtain scores on a construct, it is important for researchers to report the reliability and validity of the scale or test scores in the sample of analysis. In multilevel research, a construct may be defined as an individual construct, a within-cluster construct, a configural (or contextual) construct, a shared (or climate) construct, or a combination of shared and configural construct (Marsh et al., 2012; Stapleton et al., 2016). Researchers should report the reliability and validity of scores based on the types of the construct.

In Example 1, we may use students' IRT-based scores on the reading assessment as a dependent variable, which represents an individual construct. To report the content validity of the IRT-based scores, we can describe the content categories and the percentage of items in each category in the reading assessment. In the third-grade assessment, for example, 20% of the items were in the basic skills and vocabulary category (e.g., letter identification), 30% in the locate/recall category (e.g., recalling information from a reading passage), 35% in the integrate/interpret category, and 15% in the critique/evaluate category. To show the construct validity of the IRT-based scores (i.e., the unidimensional assumption underlying the IRT model), researchers could report the principal component analyses (PCA) results (Najarian et al., 2019). To show the reliability of the IRT-based scores, researchers could report the reliability coefficient based on the ratio of the error variance to the total variance. For example, the reliability coefficient was 0.86 for the third-grade IRT-based reading scores, which was typical and adequate for a test with 24 items (Najarian, et al., 2019).

For parent involvement in Example 1, if a researcher examines the within-school effect of parent involvement (i.e., group mean centered parent involvement) and the contextual effect of parent involvement (e.g., school mean parent involvement) separately, the psychometric characteristics of the two predictors should be reported separately. Because the scale for parent involvement consists of multiple items (e.g., attending an open house or back-to-school night; volunteering at the school or serving on a committee), researchers should report the intra-class correlations (ICCs)[1] of the

[1] There are two types of ICCs: ICC1=B2B2+W2 and ICC2=B2B2+(W2n) where B2 is the between-cluster variance, W2 is the within-cluster variance, and n is the average cluster size.

items and the composite scores as preliminary evidence for the reliability of the aggregated school means. If the ICCs are non-trivial, a multilevel CFA (MCFA) can be applied to examine the construct validity of the within-cluster composite scores and the cluster mean composite scores (see Figure 5 in Stapleton et al., 2016). Cross-level invariance (i.e., equal factor loadings at the within-cluster and between-cluster level) is required for the construct validity of a configural variable. Hence, it is important to test the cross-level invariance and report the test results (Jak et al., 2013). If cross-level invariance holds, researchers should report factor intra-class correlations (ICCs) to show the variability of the construct across clusters (Kim et al., 2016). On the other hand, if cross-level invariance does not hold, it is inappropriate to use the cluster mean centered parent involvement composite scores or the cluster mean composite scores. Instead, factor scores estimated at each level based on the MCFA could be used. For model fit evaluation, researchers should report level-specific model fit indexes (e.g., level-specific RMSEA, level-specific CFI, etc.), because the commonly used overall model fit indexes (e.g., overall RMSEA, overall CFI) are dominated by the within-cluster model fit and insensitive to misspecifications of the between-cluster model (Hsu et al., 2015; Ryu & West, 2009). When the adequate model fit is achieved, researchers could assess and report level-specific composite reliability (ω^B and ω^W) according to Lai (2020).

Readers are referred to Stapleton et al. (2016) and Stapleton and Johnson (2019) for MCFA models to evaluate the construct validity of a shared construct and a simultaneous share and configural construct. The formulas for computing level-specific composite reliabilities for these types of constructs are available in Lai (2020).

9.6 Missing data Treatment

Missing data are commonly encountered in applied research. To understand the prevalence and nature of missing data, and the impact of missing data on statistical results, researchers should report the frequency or percentages of missing data, the empirical evidence and/or theoretical arguments for the causes of missing data, and the methods employed for addressing missing data (Appelbaum et al., 2018). However, the review by Luo et al. (2021) showed that more than 70% of the studies did not report the missing mechanism and about 30% of studies did not discuss how missing data were handled in the analyses.

In multilevel research, the frequency or percentages of missing data should be reported for each variable at each level separately. In Example 1, researchers should report the percentage of students who had missing data on the dependent variable (i.e., IRT-based scores on the reading assessment) and the percentage of missingness on parent involvement composite scores. In Example 2, researchers should report the percentage of students who had self-control scores at all time points (i.e., complete data), missing scores at the first time point, the second time point, etc.

After reporting the prevalence of missing data, researchers should examine the potential mechanism for missing data, such as missing completely at random

(MCAR), missing at random (MAR), or missing not at random (MNAR). According to Rubin's (1976) missing data theory, data are MCAR if missingness is unrelated to the data, observed or missing; MAR occurs when missingness is only related to observed data; and MNAR occurs when missingness is related to missing data themselves. Data sets are likely to be MCAR when missingness is controlled by design (e.g., one or more variables are observed only on a random subsample from an initial sample). Researchers could test whether participants with missing data on a given variable, say *y*, and those with complete data on *y*, have the same distribution of the remaining variables of interest, by using Little's (1988) multivariate test, univariate t-tests for continuous variables, or chi-square test for categorical variables. Although the absence of differences in the distributions does not fulfill the conditions of MCAR, the presence of differences does rule out MCAR (Raykov, 2011). In Example 1, we could test whether students with missing data on parent involvement had different mean reading scores compared to those without missing data. If there is a difference in the means, then the data are not MCAR. Similarly, in Example 2 we could test if students who had low self-control scores at the first time point are more likely to have missing data at the second time point. If there is a relationship, then the data are not MCAR. It should be noted that even if there is no difference in the means in Example 1 or no relationship in Example 2, we still cannot confirm MCAR. In fact, it is difficult to determine which mechanism applies to missing data in practice because the determination of the missing data mechanism requires knowledge of the missing data themselves. Readers are referred to Enders (2010) and Graham (2012) for more information on missing data analysis.

Because in most cases we can only rule out MCAR empirically but not confirm the missing mechanism, it is recommended that we use modern missing data handling methods such as multiple imputation (MI) and full information maximum likelihood (FIML) estimation methods which only require MAR, instead of using listwise deletion which requires a more restrictive MCAR assumption (e.g., Enders, 2017). FIML[2] for multilevel models with missing data is often limited because most software can only handle missing outcome variables. On the other hand, MI can handle missing data on both predictors and outcome variables. In Example 1, if there are missing data on both student's reading scores and parent involvement scores, then MI is a better choice. In Example 2, because missing data are only present in the dependent variable (i.e., self-control) but not in the predictor (i.e., time), FIML is well suited.

MI includes three phases: the imputation phase, the analysis phase, and the pooling phase. For the imputation phase, researchers should report the imputation models and auxiliary variables used to generate plausible values for the missing data. There are two principal approaches for the imputation phase: Joint modeling (JM) and fully conditional specification (FCS) (e.g., Enders et al., 2016). JM is preferred when the analysis model is a random intercept model, whereas FCS is advantageous when the analysis model includes random slopes and cross-level interactions (e.g., Enders

[2] FIML includes full information maximum likelihood and full information restricted maximum likelihood estimation methods.

et al., 2018; García-Patos & Olmos, 2020). Researchers should also report the number of imputed datasets and the statistical software used for the imputation (e.g., the *jomoImpute* function in the *mitml* package for JM, or the *mice* package for FCS). For the analysis phase, researchers should show the specification of the analysis model (see the following section of Model Specification for details). For the pooling phase, researchers need to apply the appropriate rules for pooling the results (e.g., Carpenter & Kenward, 2013; Rubin, 1987) and report the final pooled results.

9.7 Model Specifications

Due to the complexity of MLMs, it is crucial to provide details on how a model is specified, including the link function used for the dependent variable, how predictors were included in the model, and how the variance structure of the random effects and errors were specified.

When an outcome is a continuous variable, the identity link function is typically used in which the outcome is directly modeled as a linear combination of predictors and random effects. Such a model is often referred to as a linear mixed model (LMM). When an outcome is a binary variable, a logit or probit link is commonly used. Such models are often called multilevel logistic (or probit) regression models. When the outcome is an ordinal variable with more than two categories, multilevel proportional odds models can be used in which a logit link function is used to transform the cumulative probability of a response being scored in a certain category or below. For count data, it is common to use a natural logarithm link function. In Example 1, students' reading IRT score can be considered as a continuous outcome, thus, LMM can be specified. In other cases, instead of using students' reading IRT score as an outcome, researchers could categorize students into a "learning difficulty" group and a "non-difficulty" group based on certain cutoff criteria (e.g., researchers have used the bottom 10% of the distribution to identify children experiencing learning difficulties and disabilities in several empirical studies; Morgan et al., 2017). In this case, a logit link function could be used to link the binary outcome (i.e., experiencing learning difficulty or not) and the predictors.

Next, researchers should document how covariates were selected and included in the models. Usually, the selection of covariates, including main predictors of interest and control variables, should be guided by research questions developed based on theories. In a case where covariates are selected using a data-driven approach, researchers should report the criteria for determining whether a covariate should be included or not. For example, when investigating the growth of self-control over time (Example 2), substantive theories might not be accurate enough for researchers to hypothesize a specific form of the growth trajectory. In this case, researchers may compare the model that specifies a linear term relative to the model that specifies both a linear and a quadratic term using the goodness of fit indices (e.g., AIC, BIC, or likelihood ratio test) to determine the shape of the growth trajectory. If covariates are centered, researchers should report the specific centering methods used.

Commonly used centering methods include group-mean centering (also known as centering within context) and grand-mean centering.

After determining the covariates to be included in a model, researchers should consider and report the specification of random effects and their covariance structure. Random effects include random intercepts and random slopes. Researchers should clearly state whether a random intercept is included in a tested model, at which level the random intercept is allowed to randomly vary, whether a lower-level covariate's slope is random, at which level the random slope is allowed to randomly vary. When there are multiple random effects at a higher level, researchers should also state how the covariance structure of the random effects is specified. In Example 1, to examine the association between student's reading IRT scores and the within-school parental involvement as well as the school mean parental involvement, a random coefficients model could be specified as shown in Equation (9.1):

$$\begin{aligned} IRTscore_{ij} &= \beta_0 + \beta_1\left(PI_{ij} - \underline{PI}_j\right) + \beta_2\underline{PI}_j + u_{0j} + \left(PI_{ij} - \underline{PI}_j\right)u_{1j} + e_{ij} \\ cov\left[u_{0j}u_{1j}\right] &= \left[\tau_{00}\tau_{10}\tau_{01}\tau_{11}\right], \mathrm{var}\left(\mathrm{e_{ij}}\right) = \sigma^2 \end{aligned} \tag{9.1}$$

where $IRTscore_{ij}$ represents the student's reading score of student *i* in school *j*, PI_{ij} represents individual student's parent involvement score, $\underline{PI}_j$ represents the mean parent involvement score in school *j*, β_0 indicates the mean intercept, β_1 indicates the mean slope of the group-mean-centered parent involvement (i.e., the within-school effect of parent involvement), β_2 indicates the contextual effect of school mean parent involvement, u_{0j} represents the random effect of school *j* on the intercept, u_{1j} represents the random effect of school *j* on the slope of the group-mean-centered parent involvement, and e_{ij} represents the student-level error term. As a part of the model, the specification of the covariance structure of the random effects should be stated. As shown in equation (9.1), an unstructured covariance matrix was adopted in this example allowing the random intercept and the random slope (u_{0j} and u_{1j}) to covary, because there was no theory or prior knowledge indicating that the two random effects were independent of each other. However, if researchers have strong rationales for the independence of the random effects, the covariance between u_{0j} and u_{1j} could be set to zero and should be reported accordingly. Furthermore, if researchers expect that the within-school effect of parent involvement would have little variation across schools, then the random effect u_{1j} could be removed from the model. In both cases, it is important for researchers to clearly state the rationales for such a priori decisions about the covariance structures of the random effects. On the other hand, if a data-driven approach is used to select the best fitting covariance structure (e.g., Matuschek et al., 2017), the procedure of model selection should be explicitly described. Finally, the covariance structure of the error term (e_{ij}) is specified as the basic identity structure indicating that the errors are independent of each other and have equal variance (σ^2) in this example. It should be noted that the error covariance structure could be more complex. For example, in models for longitudinal data, the errors could be correlated. In Example 2, we could specify an auto-regressive structure for the errors.

In addition, if the outcome is not a continuous variable that requires a non-identity link function to be used, a generalized linear mixed effects model (GLMM) should be used. Equation 9.2 shows a two-level logistic regression model (i.e., a special case of GLMMs) in Example 1 when the outcome is changed to students' status of experiencing learning difficulty (a dichotomous variable):

$$\begin{aligned} &Learningdifficulty_{ij}\ Binomial\left(1, \varphi_{ij}\right) \\ &logit\left(\varphi_{ij}\right) = \beta_0 + \beta_1\left(PI_{ij} - \underline{PI}_j\right) + \beta_2\underline{PI}_j + u_{0j} + \left(PI_{ij} - \underline{PI}_j\right)u_{1j} \\ &cov\left[u_{0j}u_{1j}\right] = \left[\tau_{00}\tau_{10}\tau_{01}\tau_{11}\right] \end{aligned} \tag{9.2}$$

As the dependent variable is in a log-scale, a nonlinear link function is adopted in Equation 9.2. φ_{ij} is the probability that a student experiences learning difficulty.β_0, β_1, and β_2 have different interpretations as opposed to their counterparts in Equation 9.1. These differences will be illustrated in the following sections.

Although it is possible to describe all the elements of the model verbally, using equations is an effective way to communicate the specified model. Thus, we recommend researchers presenting models in equations if space allows or in supplemental materials, especially when a model involves several covariates, several random effects, and/or cross-level interactions. Researchers should pay special attention to the communication of the covariance structure of the random effects. As shown in the review by Luo et al. (2021), more than 70% of the studies did not report whether there is a covariance between a random intercept and a random slope.

9.8 Estimation Methods and Software

Reporting of the estimation method and software program is important for a few reasons. First, different estimation methods may lead to different results. In order for other researchers to replicate the results, details about the specific estimation method used need to be reported. Second, each estimation method has its own pros and cons. It is important for researchers to provide the details so that the appropriateness of the method can be evaluated in the peer-review process. However, it was found that although software programs used in the studies were well documented, estimation methods and algorithms were often unreported in current practice (Luo et al., 2021). Below we provide a brief review of the commonly used estimation methods in LMMs and GLMMs and the corresponding software.

For LMMs, the maximum likelihood (ML) method is one of the most commonly used estimation methods including full maximum likelihood (FML) and restricted maximum likelihood (REML). REML has the advantage over FML for small sample sizes because REML produces more accurate estimates for variance components and standard error estimates for fixed effects (Snijders & Bosker, 2012; Lindstrom &

Bates, 1988). These estimation methods and associated approximation methods for small sample sizes can be found in the commonly used computer programs (see Luo et al., 2021 for a review).

For GLMMs, approximation methods are commonly used, including linearization methods (model approximation) and integral approximation methods. The linearization approaches create a normalized pseudo-outcome so that model can be estimated with traditional methods for LMMs (Stroup, 2012). There are two similar linearization methods: penalized quasi-likelihood (PQL) (Breslow & Clayton, 1993; Schall, 1991) and pseudo-likelihood (PL) (Wolfinger& O'Connell, 1993). The linearization methods are quite flexible in specifying residual covariance structure and can handle models with many random effects or crossed random effects (Luo et al., 2021; McNeish, 2016). Nonetheless, they would yield a quasi-likelihood rather than a true likelihood because each linearization update depends on the current pseudo-data. Hence, we can neither perform likelihood ratio tests (LRTs) among nested models nor compute commonly used likelihood-based fit statistics such as deviance, AIC, and BIC. PQL is available in SAS Glimmix (SAS Institute, 2017), and also available in R package *lme4* (Bates et al., 2015). PL is the default estimation method in SAS *Glimmix*.

Unlike linearization methods, integral approximation maximizes the actual likelihood function and thus yields the true likelihood rather than a quasi-likelihood. Two commonly used methods are Adaptive-Gauss-Hermite quadrature *(*AGQ*)* and Laplace approximation. AGQ is available in Stata (StataCorp, 2019), SAS (SAS Institute, 2017), the R package *lme*4, and M*plus* (Muthén & Muthén, 1998–2017).The Laplace approximation is available in R packages *lme4*, *glmmADMB* (Fournier et al., 2012) as well as *glmmTMB* (Brooks et al., 2017), and also available in Stata and SAS.

Bayesian estimation is an alternative estimation method characterized by its flexibility to handle multilevel data with complex variance structures (e.g., Baldwin & Fellingham, 2013), ability to deal with very small sample sizes (e.g., McNeish & Stapleton, 2016) and to incorporate existing information about a research area. Interested readers are referred to Baldwin and Fellingham (2013) for examples of how to select plausible and thoughtful prior distributions for MLMs and Zondervan-Zwijnenburg et al. (2017) for creating informative prior distributions with small sample sizes. Bayesian estimation can be conducted in MLM programs such as MLwiN (Browne, 2019) and M*plus*, R packages *MCMCglmm* (Hadfield, 2010) and *brms* (Bürkner, 2017), as well as general purpose Bayesian modeling programs such as SAS PROC MCMC, OpenBUGS (or WinBUGS; Spiegelhalter et al., 2014), and R package *INLA* (Rue et al., 2009). As statistical programs use different default priors that are not necessarily the most appropriate priors, we caution that the inconsistent results can be obtained by using default priors with different programs. Therefore, it is important for empirical researchers to report priors adopted under the investigation.

In summary, researchers should choose the appropriate estimation method based on the model and the data characteristics as well as provide sufficient information for peer review and replication purposes. In most cases, information on the estimation method and software can be communicated effectively in one sentence in the description of the data analysis. For more complex or less commonly used estimation methods (e.g., Bayesian approach), more details including software codes should be provided in a technical appendix.

9.9 Statistical Inference

Similar to the estimation methods, there are several methods for statistical inferences in MLM which have different statistical properties, thus it is important for applied researchers to report the specific statistical tests used for peer review and replication purposes. Below we provide a brief review of commonly used statistical test for fixed effects and variance components in MLMs.

To determine the statistical significance of a fixed effect in LMMs, there are three commonly used tests when using the ML estimation method: the Wald test (or the z test for a single parameter), the t test or F-test, and the LRT. When the number of clusters is large, these tests would yield similar results. However, when the number of clusters is small, Wald test and LRT are anticonservative (Halekoh & Højsgaard, 2014; Luke, 2017), while t test or F-test with small sample approximations provide more accurate estimates for standard error and/or degrees of freedom. Such approximation methods as Satterthwaite approximation and Kenward-Roger approximation can be found in SAS Mixed, SPSS, Stata mixed or xtmixed, as well as R packages *pbkrtest* (Halekoh & Højsgaard, 2014) and *lmerTest* (Kuznetsova et al., 2017).

The choice of a significance test for a fixed effect in GLMMs depends on the estimation methods. For GLMMs estimated by PQL or PL, only Wald and F or t tests can be conducted. In contrast, when GLMMs are estimated by integral approximation, Wald test, F or t test, and LRT are all available. PL has a residualized form (RPL) analogous to REML, which allows researchers to adopt the t test or F test with ad-hoc version of Kenward-Roger (Kenward & Roger, 1997) or Satterthwaite approximation (Satterthwaite, 1946). PL with approximation methods can be implemented in SAS Glimmix.

Although significance tests for the fixed effects were well reported, confidence intervals (CIs) were seldomly reported (Luo et al, 2021). Reporting CIs is an important practice because CIs provide additional information on how big or small the true effect or relationship might plausibly be. Confidence intervals for fixed effects can be obtained using either the Wald test approach or the t-test approach based on the estimated standard error and/or degrees of freedom. As an alternative, the profile likelihood confidence interval (PLCI) based on the asymptotic chi-square distribution can be used in lieu of the Wald or t type intervals. Currently, PLCI can be obtained by using *confint* function in the R package *lme4* and SAS Glimmix.

For variance components, researchers report much less information in practice, probably because variance components are usually not of the main interest (Luo, et al, 2021). However, variance component estimates can provide important information regarding the variability of an effect across clusters, hence it is important to report the basic information of variance component estimates (e.g., point estimates, standard error estimates, significance test, and confidence intervals). Researchers should be aware that the traditional Wald tests are not appropriate for variance components as the sampling distribution of the variance components is positively skewed especially when the parameter values are close to 0. A better choice is the restricted likelihood ratio test (RLRT) which compares a model that contains the tested variance component with a reference model that excludes it (Berkhof & Snijders, 2001; Chen et al., 2019).

In summary, researchers should state the specific type of statistical test chosen for testing a parameter in an MLM with the possible rationale behind the choice. Software codes should be provided in appendix or supplementary materials for transparency.

9.10 Interpretation of Regression Coefficients

The interpretation of regression coefficients in MLMs is similar to that in single-level regression models. The only difference that researchers should pay attention to is the interpretation of regression coefficients in GLMMs which represent a *conditional* effect holding other predictors and clusters fixed rather than an *average* (or marginal) effect across all clusters. For example, in the two-level logistic regression model as shown in Equation 9.2, when interpreting the regression coefficient β_1, we should obtain the exponential of β_1 [i.e., $exp(\beta_1)$], and interpret $exp(\beta_1)$ as the *multiplicative* effect of the group-mean-centered parent involvement on the odds of having learning difficulty for students *in a given school*. It is inappropriate to interpret $exp(\beta_1)$ as the *average* multiplicative effect of the group-mean-centered parent involvement *across all schools*.

It is noted that some statistical software report both conditional (subject-specific) and marginal (population-averaged) effects in the output (e.g., HLM software) while others report either conditional or marginal effects depending on the choice of the estimation methods (e.g., SAS *Glimmix: METHOD = RMPL* represents a restricted version of marginal pseudo likelihood which produce estimates of marginal effects; *METHOD = RSPL* represents a restricted version of subject-specific pseudo-likelihood which produce estimates of conditional effects). Therefore, it is important for researchers to be aware of these differences, use the appropriate type of estimates based on their research questions, and interpret the effects accordingly.

9.11 Effect Sizes

Although effect sizes are important statistics that show quantitatively the magnitude of an effect of interest and are widely used in meta-analyses, researchers using MLMs often fail to report effect sizes (Luo et al, 2021). Below we briefly introduce two types of effect sizes in the context of MLMs: standardized mean differences and proportion of variance explained.

In single-level studies, the standardized mean difference is commonly used in randomized control trials or quasi-experimental designs and is defined as the mean difference between two groups standardized with respect to the pooled standard deviation of the outcome variable. In two-level studies (e.g., cluster-randomized trials), the pooled total variance of an outcome (σ_T^2) is partitioned into a between-cluster component (σ_B^2) and a within-cluster component (σ_W^2) such that $\sigma_T^2 = \sigma_B^2 + \sigma_W^2$. Thus, there are three effect size parameters δ_T, δ_B, and δ_W in a two-level study with respect to standard deviations σ_T, σ_B, and σ_W, respectively (Hedges, 2007). As Hedges demonstrated, δ_T is the most comparable effect size with those in multisite studies or those in studies that use an individual rather than a cluster assignment strategy in a meta-analysis; δ_B is of interest in cases where treatment effect is defined in the cluster level or cluster means are used as the unit of analysis in a meta-analysis; δ_W might be the effect size most comparable with those in single-site studies in a meta-analysis.

Unlike standardized mean differences which can only be used when a predictor is a categorical variable, the proportion of variance explained (i.e., R^2 measures) can be used for both categorical and continuous predictors. In MLMs, the definition of R^2 measures is complicated due to the unexplained variances in different levels and the different sources of explained variance. Several R^2 measures are proposed in previous work (e.g., Bryk & Raudenbush, 1992; Nakagawa & Schielzeth, 2013; Snijders& Bosker, 1994; Xu, 2003). However, there is no unified framework of different R^2 measures in MLMs until a recent study by Rights and Sterba (2019). They showed a full partitioning of the outcome variance and defined twelve meaningful R^2 measures. The authors recommend researchers to report multiple R^2 measures to show a complete picture of how a model can explain variance. These R^2 measures can be obtained in the R package *r2mlm.*

9.12 Assumption Checking

MLMs have certain assumptions that need to be satisfied in order to obtain valid statistical results. Researchers should report the methods used for assumption checking and the results. In the practice of MLM, the process of checking assumptions has always been rarely reported in empirical studies (Dedrick et al, 2009; Luo et al., 2021). In the case of assumption violations, researchers need to report the measures taken to avoid potential biases. Below we provide a brief review of the assumptions underlying MLMs.

For typical LMMs, we assume that the level 1 errors and higher-level random effects are independently and normally distributed with homogeneous variance. When the number of clusters is large (e.g., >100), the violation of the normality assumption has little impact on the estimation of fixed effects and variance components (McCulloch & Neuhaus, 2011). However, the statistical inference will be impaired when the normality assumption is violated with a relatively small number of clusters (e.g., 50–100; Seco et al., 2013) and data missing at random (Lu et al., 2009).

For GLMMs, only the higher-level random effects are assumed to follow a normal distribution with homogenous variance. In addition, it is assumed that the conditional mean and variance of the outcome have a theoretical relationship as the specified distribution (e.g., Poisson or binomial distribution) suggested. Violation of this assumption would lead to overdispersion, causing a poor fit to the data and misleading inferences (Hilbe, 2011, 2014). Computer programs for MLM diagnostic include the R package *HLMdiag* (Loy, 2016) and a Stata (StataCorp, 2019) post estimation package "MLT" (Moehring & Schmidt, 2013). Researchers are also recommended to conduct sensitivity analysis by using both standard estimation and a robust approach in the case of assumption violations (Agresti et al., 2004), such as bootstrap and rank-based methods.

9.13 Discussion

The discussion part allows researchers to interpret the results and justify the significance of the current study. The discussion should be closely related to the research purposes and original questions, support their arguments by linking the findings to previous studies, and illustrate the theoretical and/or practical implications. Researchers should also provide the limitations of the current study due to flaws in designs, sampling procedures, measurements, and statistical analyses. In the final section, researchers can provide possible future directions and extensions based on the current findings and previous work. In addition to the above general guidelines, researchers can discuss the new information provided by using MLMs compared with the traditional regression analysis. Besides, a discussion about how the results are impacted by certain decisions in the data analysis process (e.g., missing data treatment, choice of covariates, model specification, etc.) will provide more insights for readers to interpret and use the findings.

9.14 Summary and Checklist

MLM is a very flexible technique for analyzing clustered data, but its complexity requires applied researchers to pay more attention to the details. It is our hope that this chapter provides useful guidelines for researchers to present their MLM applications in a clear and rigorous way with all the necessary details for transparency and replicability. We understand that not all the technical details can be presented in a paper itself due to the limit of space, however, with the increasing availability of online publications, those technical details could be provided in online supplemental materials for interested readers. Finally, we provide a checklist to describe the main points that should be thoroughly checked and clearly conveyed in reporting MLM applications.

- Describe the purpose of a study and clearly state all the research questions, differentiating primary vs. secondary research questions, a priori vs. post hoc questions, and confirmatory vs. exploratory questions.
- Report whether the sample is a probability or a nonprobability sample and if a probability sample is obtained, what sampling methods are employed to achieve the probability sample.
- Explicitly report whether sampling weights are employed or not, and if multiple sets of sampling weights are used, researchers should be clear about which sampling weights are applied to which level or which variables in each level.
- Report the total sample size at each level as well as the variation of cluster size across clusters.
- Provide an assessment as to whether the obtained sample size is adequate for estimating and testing the effects of interest, preferably by a power analysis rather than relying on the general sample size guidelines.
- Report the reliability and validity of the scale or test scores in the sample of analysis according to the type of a construct.
- Report the frequency or percentages of missing data, the empirical evidence and/or theoretical arguments for the causes of missing data, and the methods employed for addressing missing data.
- Provide details on how a model is specified, including the link function used for the dependent variable, how predictors are included in the model, and how the variance structure of the random effects and errors are specified.
- Report the estimation method and software program
- Report the point estimates, standard error estimates, significance test results with p values, and confidence intervals for both fixed effects and variance components.
- Report the specific statistical tests used for statistical inferences of fixed effects and variance components. For tests that involve approximation methods for degrees of freedom, the specific method used should be reported.
- Interpret the regression coefficients appropriately, especially when a GLMM with non-identity link is used.
- Report effect sizes for the effects of interest.

- Report assumption checking results. In the case of assumption violations, report the measures taken to avoid the potential biases.
- Provide explanations of the current findings, illustrate the theoretical and/or practical implications, discuss how the results are impacted by certain data analysis decisions (e.g., missing data treatment, choice of covariates, model specification, etc.).

References

Agresti, A., Caffo, B., & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics &amp; Data Analysis, 47*(3), 639–653. https://doi.org/10.1016/j.csda.2003.12.009

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *AmericanPsychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics-Theory and Methods, 35*, 439–460. https://doi.org/10.1080/03610920500476598

Baek, E., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2020). Brief research report: Bayesian versus REML estimation with noninformative priors in multilevel single-case data. *Journal of Experimental Education, 88*, 698–710. https://doi.org/10.1080/00220973.2018.1527280

Baek, E., & Ferron, J. M. (2020). Modeling heterogeneity of the level-1 error covariance matrix in multilevel models for single-case data. *Methodology, 16*, 166–185. https://doi.org/10.5964/meth.2817

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18*(2), 151–164. https://doi.org/10.1037/a0030642

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9–25. https://doi.org/10.2307/2290687

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal, 9*(2), 378–400. https://doi.org/10.3929/ethz-b-000240890

Browne, W. J. (2019). *MCMC estimation in MLwiN Version 3.03.* Centre for Multilevel Modelling, University of Bristol.

Berkhof, J., & Snijders, T. A. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, *26*(2), 133–152. https://doi.org/10.3102/10769986026002133

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Sage.

Bürkner, P. C. (2017). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10/1*(July), 395–411. https://doi.org/10.32614/RJ-2018-017

Carpenter, J., & Kenward, M. (2013). *Multiple imputation and its application.* Wiley.

Chen, S. T. (2019). *glmmVCtest: Testing variance components in generalized linear mixed models.* R package version 0.1.0.

Chen, S. T., Xiao, L., & Staicu, A. M. (2019). *An approximate restricted likelihood ratio test for variance components in generalized linear mixed models* [Manuscript submitted for publication].

Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B: Statistical Methodology, 66*(1), 165–185. https://doi.org/10.1111/j.1467-9868.2004.00438.x

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research, 79*(1), 69–102. https://doi.org/10.3102/0034654308325581

Dong, N., & Maynard, R. A. (2013). PowerUp! A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasiexperimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143

Donohue, M. C., Gamst, A. C., Edland, S. D., & Donohue, M. (2013). Package "longpower." *Biometrics, 53*, 937–947.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy, 98*, 4–18. https://doi.org/10.1016/j.brat.2016.11.008

Enders, C. K., Hayes, T., & Du, H. (2018). A comparison of multilevel imputation schemes for random coefficient models: Fully conditional specification and joint model imputation with random covariance matrices. *Multivariate Behavioral Research, 53*(5), 695–713. https://doi.org/10.1080/00273171.2018.1477040

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods, 21*(2), 222–240. https://doi.org/10.1037/met0000063

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*, 372–384. https://doi.org/10.3758/BRM.41.2.372

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell & D. Betsy Mccoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Information Age Publishing.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., & Sibert, J. (2012). AD Model Builder: Using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods &amp; Software, 27*(2), 233–249. https://doi.org/10.1080/10556788.2011.597854

García-Patos, P., & Olmos, R. (2020). Multiple imputation in multilevel models. A revision of the current software and usage examples for researchers. *The Spanish Journal of Psychology, 23*, E46. https://doi.org/10.1017/SJP.2020.48

Gastanaga, V. M., McLaren, C. E., & Delfino, R. J. (2006). Power calculations for generalized linear models in observational longitudinal studies: A simulation approach in SAS. *Computer Methods and Programs in Biomedicine, 84*, 27–33. https://doi.org/10.1016/j.cmpb.2006.07.011

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software, 33*(2), 1–22. https://doi.org/10.18637/jss.v033.i02

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models: The R package pbkrtest. *Journal of Statistical Software, 59*(9), 1–30. https://doi.org/10.18637/jss.v059.i09

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data.* Cambridge University Press.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer.

Hsu, H. Y., Kwok, O. M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research, 50*(2), 197–215. https://doi.org/10.1080/00273171.2014.977429

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(2), 265–282. https://doi.org/10.1080/10705511.2013.769392

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983–997. https://doi.org/10.2307/2533558

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research, 51*(6), 881–898. https://doi.org/10.1080/00273171.2016.1228042

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies.* California State University.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lai, M. H. C. (2020). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods, 26*(1), 90–102. https://doi.org/10.1037/met0000287

Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association, 83*(404), 1014–1022. https://doi.org/10.1080/01621459.1988.10478693

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Loy, A. (2016). *Package HLMdiag* (Version 0.4.0) [Computer software]. http://cran.rproject.org/web/packages/HLMdiag/HLMdiag.pdf

Lu, N., Tang, W., He, H., Yu, Q., Crits-Christoph, P., Zhang, H., & Tu, X. (2009). On the impact of parametric assumptions and robust alternatives for longitudinal data analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences, 51*(4), 627–643. https://doi.org/10.1002/bimj.200800186

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting Practice in Multilevel Modeling: A Revisit After 10 Years. *Review of Educational Research, 91*(3), 311–355. https://doi.org/10.3102/0034654321991229

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97*(5), 951–966. https://doi.org/10.1037/a0028380

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science, 26*(3), 388–402. https://doi.org/10.1214/11-STS361

McNeish, D. (2016). Estimation methods for mixed logistic models with few clusters. *Multivariate Behavioral Research, 51*(6), 790–804. https://doi.org/10.1080/00273171.2016.1236237

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research, 52*(5), 661–670. https://doi.org/10.1080/00273171.2017.1344538

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

Moehring, K., & Schmidt, A. (2013). *"MLT": Module providing different tools for multilevel modeling* [Computer software]. Boston College Department of Economics.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology, 7*(1), 1–10. https://doi.org/10.1186/1471-2288-7-34

Morgan, P. L., Li, H., Farkas, G., Cook, M., Pun, W. H., & Hillemeier, M. M. (2017). Executive functioning deficits increase kindergarten children's risk for reading and mathematics difficulties in first grade. *Contemporary Educational Psychology, 50*, 23–32. https://doi.org/10.1016/j.cedpsych.2016.01.004

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th edition). Muthén & Muthén.

Najarian, M., Tourangeau, K., Nord, C., Wallner-Allen, K., & Vaden-Kiernan, N. (2019). Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K: 2011), Third-Grade, Fourth-Grade, and Fifth-Grade Psychometric Report (NCES 2020-123). U.S. Department of Education. Institute of Education Sciences. Washington, DC: National Center for Education Statistics. Retrieved June 17, 2021 from https://nces.ed.gov/pubsearch.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*(2), 317–337. https://doi.org/10.2307/1403631

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology, 60,* 23–56. https://doi.org/10.1111/1467-9868.00106

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173–185. https://doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*, 199–213. https://doi.org/10.1037/1082-989X.5.2.199

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X. F., Martinez, A., & Bloom, H. (2011). *Optimal design software for multi-level and longitudinal research* (Version 3.01) [Computer software]. William T. Grant Foundation. http://www.wtgrantfoundation.org

Raykov, T. (2011). On testability of missing data mechanisms in incomplete data sets. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(3), 419–429. https://doi.org/10.1080/10705511.2011.582396

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods, 24*(3), 309–338. https://doi.org/10.1037/met0000184

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical Methodology), 71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling, 16*(4), 583–601. https://doi.org/10.1080/10705510903203466

SAS Institute. (2017). *SAS/STAT® 14.3 user's guide.*

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*(6), 110–114. https://doi.org/10.2307/3002019

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika, 78*(4), 719–727. https://doi.org/10.1093/biomet/78.4.719

Seco, G. V., García, M. A., García, M. P. F., & Rojas, P. E. L. (2013). Multilevel bootstrap analysis with assumptions violated. *Psicothema, 25*(4), 520–528. https://doi.org/10.7334/psicothema2013.58

Snijders, T. A., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods &amp; Research, 22*(3), 342–363. https://doi.org/10.1177/0049124194022003004

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd edition). Sage.

Snijders, T. A., Bosker, R. J., & Guldemond, H. (1996). *PINT user's manual* (Version 1.6) [Computer software]. KIP. https://www.kip.com/software.php

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2014). *OpenBUGS user manual* (Version 3.2.3) [Computer software manual]. http://www.openbugs.net/Manuals/Manual.html

Stapleton, L. M., & Johnson, T. L. (2019). Models to examine the validity of cluster-level factor structure using individual-level data. *Advances in Methods and Practices in Psychological Science, 2*(3), 312–329. https://doi.org/10.1177/2515245919855039

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481–520. https://doi.org/10.3102/1076998616646200

StataCorp. (2019). *Stata statistical software: Release 16*.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC Press.

Thomas, L. S., & Heck, H. R. (2001). Analysis of large-scale secondary data in higher education research: potential perils associated with complex sampling designs. *Research in Higher Education, 42*(5), 517–540. https://doi.org/10.1023/A:1011098109834

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models a pseudo likelihood approach. *Journal of Statistical Computation and Simulation, 48*(3–4), 233–243. https://doi.org/10.1080/00949659308811554

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine, 22*(22), 3527–3541. https://doi.org/10.1002/sim.1572

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & Van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development, 14*(4), 305–320. https://doi.org/10.1080/15427609.2017.1370966

Wen Luo is a Professor in the Research, Measurement, and Statistics (RMS) program in the Department of Educational Psychology at Texas A&M University. Her research focuses on developing, evaluating, and applying quantitative modeling techniques, with a specialization in multilevel modeling. She has been the principal investigator or co-principal investigator in externally funded grant projects of over 1.5 million dollars. Dr. Luo is on the editorial board of a top-tier journal. She has published 39 peer-reviewed journal articles and was the Chair of the Multilevel Modeling Special Interest Group of the American Educational Research Association.

Eunkyeng Baek is an Assistant Professor in the program of Research, Measurement, and Statistics (RMS) in the Department of Educational Psychology at Texas A&M University. Her areas of expertise are multilevel modeling (MLM) using single-case experimental data (SCED), Bayesian estimation, and the applications of these techniques in educational and psychological data. She is conducting various methodological studies as well as applied studies that analyze SCED or meta-SCED data using MLM and Bayesian estimation.

Haoran Li is a doctoral student in the Research, Measurement, and Statistics (RMS) program in the Department of Educational Psychology at Texas A&M University. He is interested in multilevel modeling, structural equation modeling, and longitudinal data analysis and their applications in empirical studies.

Chapter 10
Application of Multilevel Models to International Large-Scale Student Assessment Data

Maciej Jakubowski and Tomasz Gajderowicz

Abstract This chapter discusses applications of the multilevel modeling to international large-scale student assessment (ILSA), focusing on OECD's PISA and IEA's TIMSS. Multilevel models are routinely applied to analyze these data. However, several methodological issues need to be addressed to use these models in empirical applications correctly. First, we discuss how plausible values in multilevel modeling affect estimates of fixed and random components. Second, we discuss how to consider survey weights to decompose variance and estimate separate within- and between-school effects. Third, we discuss the use of replicate weights and compare standard errors estimated with this method to those typically obtained in multilevel modeling with robust standard errors. Fourth, we discuss applications of more complex multilevel models, like three-level models and models with cross-level effects. We summarize by providing key points to consider for researchers when applying multilevel modeling with ILSA data.

Keywords International large-scale assessment · Plausible values · Survey weight · Replicate weight · Three-level model · Cross-level effect

10.1 Introduction

The educational data have a hierarchical structure as students are nested in classrooms, and classrooms are nested in schools. Thus, multilevel modeling is a natural choice for this type of data. Examples of multilevel regressions with school and student data are presented in most books discussing applications of these statistical models. Moreover, the whole structure of education systems relies on several nested layers as schools are often managed by local authorities, governed or supervised by regional or subnational entities, and finally, education systems are organized at the subnational or national level.

M. Jakubowski (✉) · T. Gajderowicz
University of Warsaw, Warszawa, Poland
e-mail: mjakubowski@uw.edu.pl

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_10

Hypotheses in empirical research in education are also often related to interactions between different governance levels. Researchers are usually interested in individual, student-level effects and between-school effects and the relationship between-country-wide policies and associations with outcomes at the local level. One can imagine adding layers related to language, culture, governance, accountability, or practices and policies. In psychometric research, models analyzing individual test or questionnaire items that are nested in students or in time periods are also applied to address issues related to measurement error.

All large-scale international assessments of students have a hierarchical structure with students nested in classrooms or schools, and then schools nested in countries. Additional layers are sometimes added when analyzing regional data, teacher effects, or item-level responses of students. Multilevel modeling with these data is popular among researchers and often involves cross-level interactions. However, important issues related to the statistical design of international large-scale assessment data need to be addressed to analyze them properly, obtain unbiased population estimates, and measure their uncertainty. This chapter discusses the usage of plausible values, survey and replicate weights, assumptions about random effects distribution, and other issues that often arise in empirical applications but are also often misunderstood or incorrectly addressed. Throughout the chapter, we provide examples using the most recent PISA and TIMSS data.

10.2 Example of Typical Use—Modeling Relationship Between Socioeconomic Background and Student Achievement in PISA 2018

There are three main advantages of multilevel modeling with large-scale student assessment data. First, they reflect the sampling structure with schools at the higher level and classrooms and students at the lower levels. We discuss below the benefits and costs of applying multilevel models to reflect the complex sampling scheme in international studies. The second advantage is that multilevel regressions provide a decomposition of the variance, and the third advantage is that they allow modeling variance at different levels, including cross-level interactions.

We use PISA 2018 data to demonstrate how to use multilevel models to analyze the relationship between student socioeconomic background and reading achievement. In education research, it is a common model used to estimate the effects net of family background. This is also a model often used in multilevel modeling textbooks, starting from the popular Raudenbush and Bryk book (2002), which opens with an example of modeling SES association with achievement. In PISA, the socioeconomic background is measured through the index of economic, social, and cultural status. This is an index that combines information about parents' education and occupation, and educational, cultural, and material resources available at home (see Avvisati,

Table 10.1 Example of multilevel analysis with PISA 2018 data—explaining reading achievement with variance decomposition and student- and school-level slopes of socioeconomic background

	(1)	(2)	(3)	(4)
	b/se	b/se	b/se	b/se
ESCS slope		16.8		
		0.66		
Between-school ESCS slope			55.1	51.6
			2.42	2.81
Within-school ESCS slope			12.5	12.5
			0.70	0.70
Constant	446.0	457.5	481.9	481.1
	2.75	2.31	1.76	2.73
Country fixed effects	No	No	No	Yes
School-level variance	5578.5	3874.4	2225.9	1904.2
	255.56	199.04	125.04	116.30
Student-level variance	6280.2	6180.8	6176.3	6175.3
	80.24	74.87	75.73	75.70
Intraclass correlation	0.47	0.39	0.26	0.24
N of schools	10,180	10,180	10,180	10,180
N of students	249,334	249,334	249,334	249,334

Note Authors' estimation with PISA 2018 microdata. Results were obtained with ten plausible values of reading achievement and a sample of all students and countries that have participated in this assessment. Student and school weights were applied with student-level weights scaled to sum to the sample size of their school

2020 for a detailed discussion of this index and its comparability across countries and Pokropek et al., 2015 for a decomposition using structural equation modeling).

Tables 10.1 and 10.2 show results for a multilevel model applied to PISA 2018 data explaining reading achievement with the ESCS index. The index was centered at the weighted school means to decompose the effects into within- and between-school effects. The estimates are based on a model with ten plausible values and different specifications regarding random components and country effects.

Table 10.1 shows results for a model with school-level random intercepts. Column (1) shows results for the empty model—the model with intercepts only. This model provides a baseline decomposition of the variance into school and student levels. One could argue that this is one of the most important findings from large-scale international assessments, that a substantial part of the total achievement variance is associated with school-level effects. In this case, for the pooled sample of all countries that participated in PISA 2018, interclass correlation shows that nearly half of the total variance is associated with schools. These estimates are obtained for a weighted sample of students with ESCS data available, so they can be used for comparisons

Table 10.2 Example of the random coefficient multilevel model with PISA 2018 data

	(1)	(2)
	b/se	b/se
Between-school ESCS slope	53.9***	55.1***
	2.30	2.43
Within-school ESCS slope	12.3***	15.3***
	0.74	0.85
Interaction between the school-average ESCS and within-school ESCS slope		5.2***
		0.96
Constant	481.1***	481.9***
	1.76	1.77
ESCS slope variance	149.2***	122.3***
	20.96	20.10
School-level variance	2250.5***	2243.2***
	125.74	126.68
Correlation (escs,constant)	0.3***	0.3***
	0.05	0.06
Student-level variance	6070.2***	6072.0***
	75.14	75.50
N of schools	10,180	10,180
N of students	249,334	249,334

Note Authors' estimation with PISA 2018 microdata. Results were obtained with ten plausible values of reading achievement and a sample of all students and countries that have participated in this assessment. Student and school weights were applied with student-level weights scaled to sum to the sample size of their school

with models incorporating ESCS effects (missing observations constituted less than 4% of the original sample).

Columns (2) and (3) compare results for an approach typical for traditional linear regression modeling with the one possible with multilevel models that decompose associations between school- and student-level effects. Column (2) shows results for a multilevel model with the ESCS index as the only explanatory variable. The slope is around 17 points meaning that one standard deviation change in the ESCS index is associated with 17 points improvement in reading scores. Column (3) shows separate estimates for between- and within-school associations of ESCS. The within-school association is slightly weaker, but the between-school association is much stronger, showing that one standard deviation increase in school-average ESCS index is associated with an improvement of more than 50 points, which is equivalent to half a standard deviation of the reading achievement distribution for OECD countries (weighting countries equally). Note also that including the school-average slope of

ESCS explains 60% of the school-level variance, while at the student level, the within-school effect explains less than 2% of the variance.

In other words, this simple model shows that socioeconomic background is a powerful predictor of average school achievement but cannot explain much of the within-school differences. While this is not a new finding in education research, international assessment data show that this relationship holds for most schools around the world. In fact, PISA data are used to compare such associations showing in which countries school composition of student socioeconomic background is a more powerful predictor of achievement and how much of the total variance is at the school level. This is a descriptive but powerful tool for comparing inequalities related to school and socioeconomic background across countries.

The last column shows a similar model but with country fixed effects. Note that the estimated coefficients for between- and within-school associations of reading achievement with the socioeconomic background are similar, so they are not driven by between-country differences in average achievement. Note also that country fixed effects are able to explain some of the school-level variance but less the school-average socioeconomic background. Again that is an interesting finding showing that differences between schools in their composition and achievement are much larger and more important than between-country achievement differences.

This model can be further expanded by slopes of explanatory variables to randomly vary and to explain this variation with, for example, cross-level interactions. Table 10.2 shows results for a model with a random coefficient for the within-school differences in the ESCS index. The model estimates the variance of intercepts at the school and student level, but also variance in the slope of the within-school ESCS index and the covariance of the ESCS slopes and school intercepts. Results show that within-school association between student ESCS and reading achievement vary significantly across schools. We hypothesize that this variation might depend on school SES composition, and this interaction is estimated by the model presented in column (2). Indeed, the higher is the average socioeconomic background of students in a school, the stronger is the relationship between within-school ESCS and achievement. This interaction effect explains around 18% of the within-school ESCS slope variation.

10.3 Applications

The above examples demonstrate the typical use of multilevel models with international large-scale student assessment data. Variance decomposition and comparisons of between- and within-school effects are commonly applied to PISA data and are similar to the first application of multilevel modeling in education (Aitkin & Longford, 1986; Raudenbush & Bryk, 1986). The approach was partly popularized by first research using PISA data (for example, Willms, 2010) and PISA OECD reports, which routinely apply these models to decompose student- and school-level relationships (see for example results presented in OECD, 2019, but also Annex A3 with notes on the technical application of these models in PISA).

The most common approach is to study school effectiveness using multilevel models with sets of school-level and student-level predictors (for a review, see Klieme, 2013). Interestingly, these models often demonstrate that learning conditions and practices at the school level are less related to achievement than student-level opinions about the teaching process. Multilevel modeling provides a unique opportunity to study this kind of question. For school-related factors, especially for studies of socioeconomic background, PISA data often provide more detailed information. For teacher-related factors, however, TIMSS and PIRLS data might be more suitable. The sampling scheme in TIMSS and PIRLS is different from whole classrooms sampled within selected schools. This opens a possibility to collect more meaningful information about teaching as questionnaires are filled by all students of a particular teacher, separately for mathematics and science. The clear link between students and their teachers opens a possibility to model this relationship with multilevel models.

The applications of multilevel modeling with PISA data go beyond typical school effectiveness research. For example, multilevel models are applied to better understand data on student wellbeing (He et al., 2019; Jakubowski & Gajderowicz, 2020; Sznitman et al., 2011), sources of bullying (Winnaar et al., 2018; Yavuz et al., 2017), and attitudes (Lu & Bolt, 2015; Pitsia et al., 2017; Sun et al., 2012). Also, the data are often combined to provide a broader picture of student achievement and related factors (for example, see Grilli et al., 2016).

The application of multilevel modeling to international student assessment data is an obvious choice, but several technical issues need to be addressed to properly estimate population relationships of interest. As we will see below, these technical issues can be addressed with a good understanding of the role of plausible values and complex sampling in deriving conclusions from multilevel models. Many statistical packages allow taking these issues into account. More complex models, for example, three-level models, are also applied to these data—however, their raise technical issues which, as discussed below, are not straightforward to address.

10.4 Plausible Values and Multilevel Models

In publicly available datasets from large-scale student assessments like PISA, TIMSS, or PIRLS, achievement results are provided as sets of the so-called plausible values. These variables reflect not only student achievement but also the uncertainty with which it is measured for the student population. Plausible values are imputations of latent student achievement. Their correct use allows obtaining unbiased estimates of achievement in student populations, correcting for measurement error when relating to other variables in standard statistical models, and obtaining proper uncertainty measures in models with student achievement (see Wu, 2005).

For some researchers, plausible values can be seen as a technical obstacle in analyzing data like PISA or TIMSS. Analysis with plausible values requires special software, commands, or the application of formulas to calculate results by hand from statistical models with separate plausible values. Thus, researchers often try to simplify the analysis with plausible values making mistakes that invalidate their results. Below we show that using plausible is quite straightforward and that common shortcuts provide highly biased results. We also show how to use plausible values to obtain initial results faster before deciding about the final model, which is often helpful in time-consuming multilevel analysis.

First, note that a single plausible value provides an unbiased point estimate. If plausible values are drawn from distributions conditional on other variables involved in the final statistical model, then correlations with single plausible values also reflect latent correlations with other variables. For example, if the final statistical model involves student gender, plausible values should be estimated based on student gender and correlation. In this case, a simple correlation between one plausible value and student gender reflects the latent correlation between gender and student achievement. Thus, estimation with one plausible value provides unbiased point estimates also for latent correlations. However, it does not capture the effect of measurement error on the estimated variance. In other words, standard errors will be downward biased as they will not reflect measurement error.

To correctly estimate point estimates and their standard errors, one needs to run separate models with each plausible value and then take the average of estimates across these models as the point estimate and use the so-called Rubin's formula to calculate their standard errors (see Rubin, 1987). A researcher can apply Rubin's formulas herself by collecting results for each plausible value and then calculating final point estimates and standard errors. Some software packages allow to use of plausible values and calculate correct results, or there are user-written packages that can do it. It is also possible to use solutions developed for multiple imputations of missing data as the formulas are the same, and correct results can be obtained after defining each plausible as an imputed variable.

Taking an intuitive uniformed shortcut by calculating first the average of all plausible values and then running statistical models with this average as a measure of student achievement is the most common mistake done by entry-level researchers. The intuition behind this approach is that the average of plausible values is still a good achievement estimate, but in reality, such a variable suffers from an artificially lower variance. Thus, depending on a model, the final results will be biased as the overall achievement variance will be underestimated, and correlations with other indicators will be overestimated (see OECD, 2009, p. 128).

Under most circumstances, it would be more advisable to use the first plausible value if it is not possible to calculate final estimates by applying Rubin's formulas to statistical models run separately with each plausible. For example, if one is mainly interested in point estimates or, for example, creates graphic illustrations of the data, using the first plausible value will suffice. Also, when exploring the data and searching for a final model, it is also advisable to use one of the plausible values to quickly run multiple models and then do proper calculations when estimating the final model. In this case, however, one should note that the results of statistical tests will be more optimistic, so with the final model, some hypotheses might be rejected even if initial findings suggest statistically significant results.

Table 10.3 illustrates the above-mentioned issues using PISA 2000 data for Poland and two-level multilevel models with students nested in schools. Models explain student reading performance but with four differently defined variables measuring achievement. The first achievement variable is the so-called Warm estimate (weighted likelihood estimate) (Warm, 1989). The results for this variable are presented in columns (1) and (5). In columns (2) and (6), results obtained with one plausible value are presented. In columns (3) and (7), the models were estimated with the average of five plausible values as the outcome variable. The columns (4) and (8) rely on Rubin's formula to calculate point estimates and standard errors from five separate multilevel models, each run with a different plausible value.

Table 10.3 Comparisons of multilevel models estimated with different measures of student achievement (PISA 2000 data for Poland and reading achievement)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Warm estimate	1^{st} PV	Mean PV	5 PVs	Warm estimate	1^{st} PV	Mean PV	5 PVs
	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se
PISA index of reading enjoyment					15.7	14.9	16.2	16.2
					(1.4)	(1.2)	(1.2)	(1.8)
Males (females as a base group)					0.4	0.5	0.8	0.7
					(2.6)	(2.4)	(2.2)	(2.5)
ISCED 3B schools (3A as a base group)					−57.2	−60.2	−61.3	−61.4
					(7.2)	(7.6)	(7.4)	(7.5)
ISCED 3C schools (3A as a base group)					−158.7	−168.6	−168.3	−168.3
					(7.8)	(8.3)	(8.0)	(8.2)
Constant	464.9	463.6	462.7	462.8	536.5	539.0	538.5	538.7
	(6.7)	(7.1)	(7.0)	(7.1)	(5.5)	(5.8)	(5.6)	(5.7)
School-level variance	5547.9	6220.1	6138.0	6135.3	996.2	1170.0	1100.8	1095.1
	(358.2)	(398.6)	(392.3)	(400.2)	(72.7)	(82.2)	(76.6)	(85.1)
Student-level variance	4438.3	3646.5	3177.4	3710.0	4167.6	3442.8	2947.6	3474.7
	(52.9)	(43.4)	(37.8)	(61.2)	(50.7)	(41.9)	(35.8)	(50.2)
Intraclass correlation	0.56	0.63	0.66	0.62	0.19	0.25	0.27	0.24
N of schools	127	127	127	127	127	127	127	127
N of students	3653	3654	3654	3654	3511	3512	3512	3512

Note own calculations using PISA 2000 data for Poland. Standard errors in parentheses

That point estimates for regression coefficients are highly similar across results with different achievement measures. The main difference lies in standard errors and in the estimates of variance components. In general, the Warm likelihood estimate will overestimate student achievement variance, while the variable calculated as the average of plausible values will underestimate it. The results with just one plausible value will provide unbiased point estimates and correct achievement variance estimates, but the standard errors will be underestimated as they do not reflect the measurement error. The results calculated using Rubin's formula should be taken as a reference point for other models. Note that by using standard error estimates from these methods, the precise value of both measurement and sampling error in the data can be found.

The so-called empty models presented in columns (1) to (4) show that the student-level variance is overestimated for the Warm measure and underestimated for the mean PV measure, as expected. The results with one or five plausible values are highly similar. For the school-level variance, the results are similar, although the estimate with the Warm likelihood measure of achievement seems to be lower, and the intraclass correlation is also lower, as it is based on an overestimated variance at the individual level. The intraclass correlation will be higher for the model with achievement measured as the average of plausible values as it underestimates individual variance. Regarding the standard errors, note the estimates for the individual-level predictors: index of reading enjoyment and gender. The estimates based on only one plausible value are the lowest as they do not reflect the measurement error.

10.5 Survey Weights Adjustments for Multilevel Models

Large-scale surveys, including student assessments, rely on complex sampling (stratification, two- or more sampling stages) and non-response adjustments. In general, the probability of sampling a student in school surveys will always vary. This is because sampling schemes always start with sampling schools first and then students (or whole classrooms) within schools. In this case, students from smaller schools are more likely to be selected than those from larger schools. As school size is usually correlated with important student and school characteristics, datasets obtained through such sampling schemes require weighting to correct for differences in sampling probabilities. Further corrections are applied to student- and school-level sampling probabilities due to non-response and oversampling of some populations (like private schools or minority-group students). These corrections vary across countries, and the correct use of survey weights is crucial for cross-country comparisons.

Without sampling weights, the results of the statistical model show estimates for the sample but are not representative of the population. However, the use of survey weights in multilevel modeling is not straightforward. Several methods are available to adjust for arising biases, but their performance will vary depending on cluster sizes, sampling schemes, the statistical model applied, and might even vary for various estimates from the same model (e.g., regression coefficients vs. variance components) (see Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006). A common piece of advice is to perform robustness checks to compare different approaches empirically in order to assure that results do not vary importantly, and if they do, to consider again assumptions made behind these corrections.

The probability weights for each sampling stage are required for multilevel models, and the multilevel model should reflect the hierarchical structure of the sampling design. Let's consider the simplest case with schools sampled first and then students sampled within schools. In this case, one should know the sampling probability for each school and calculate the weight as the inverse of this probability. One should also know the sampling probability after a student's school was selected and then calculate the conditional sampling weight as the inverse of this probability. Only the final combined sampling weight is available in most surveys, which reflects the inverse probability of being sampled without specifying probabilities at each sampling stage. In this case, one can calculate the conditional probability weight by dividing unconditional probability by school probability weight.

Even if sampling probabilities were available at each stage, the scale of weights at the lowest level (student level in our examples) affects the estimation of multilevel equations, which is different from standard approaches like linear regression, where the scale of weights is unimportant. Therefore, re-scaling of survey weights is necessary, but different methods can produce varying results, and it is unknown which is best fitted for the sampling scheme considered and for the analyzed population. Below we discuss three weight re-scaling methods, which are commonly applied in multilevel modeling of survey data.

As an empirical example, we estimate a two-level model with students nested in schools using the dataset from PISA 2018 with all OECD countries. The model explains student achievement using the PISA's economic, social, and cultural status (ESCS), which is an index combining information on parents' education, occupation, and family resources at home. This index highly correlates with student achievement in all countries, which is a typical finding for educational research. Students with disadvantaged backgrounds have on average lower achievement than students from privileged families. However, countries do differ in the extent to which socioeconomic background is related to achievement, which is often interpreted as a measure of inequality. A stronger relationship with performance shows larger differences in achievement depending on the socioeconomic background when compared to countries with a weaker relationship. Moreover, with multilevel models, it is possible to separate within- and between-school associations, which again can be used as a measure of segregation within and between schools by students' socioeconomic background.

Table 10.4 compares unweighted results with results weighted by student-level weights only, school-level weights only, and weighted with both student- and school-level weight with three different adjustment methods. The first method re-scales the student-level weights to be the sum of the cluster size. The second method re-scales the student-level weights to be the sum of the "effective" cluster size. These two methods do not re-scale the school-level weights, but the third method replaces the

Table 10.4 Comparison of unweighted and weighted multilevel models with different scaling methods of student-level weights—example using PISA 2018 data for OECD countries

	(1)	(2)	(3)	(4)	(5)	(6)
	No weights	Student final weights only	School weight only, no scaling	Student and school weight, scaling to cluster size	Student and school weight, scaling to effective cluster size	Student and school weight, GK scaling method
	b/se	b/se	b/se	b/se	b/se	b/se
ESCS index	19.3	37.9	16.9	16.8	16.8	19.3
	0.2	0.7	0.7	0.7	0.7	0.7
Constant	484.5	495.0	457.7	457.5	457.5	483.9
	0.6	1.1	2.3	2.3	2.3	1.5
School-level variance	2766.7	0.0	3860.4	3874.4	3873.2	3032.7
	46.7	0.0	198.0	199.0	199.0	133.3
Student-level variance	6721.9	9059.7	6178.1	6180.8	6178.4	6791.7
	21.1	202.5	74.7	74.9	74.8	92.9
Intraclass correlation	0.29	0.00	0.38	0.39	0.39	0.31
N of schools	10,180	10,180	10,180	10,180	10,180	10,180
N of students	249,334	249,334	249,334	249,334	249,334	249,334

Note All models are estimated with ten plausible values of reading achievement; dataset includes all OECD countries that have participated in PISA 2018 and have available reading achievement data

weights at the school level with the cluster averages of the combined student-level survey weights (the product of school weight and the conditional student weight) and then sets student-level weights to 1 (for detailed formulas and estimation methods see Graubard & Korn, 1996; Rabe-Hesketh & Skrondal, 2006).

The results presented in Table 10.2 demonstrate that the weighting and scaling of student-level weights play an essential role in interpreting results from the multilevel models. Column (1) shows a model without survey weights. This model shows estimates for the sample of students from OECD countries. It has no interpretation in terms of relationships in the population of OECD students. Model in column (2) uses student weights only, for which coefficients for fixed effects of ESCS index and constant are identical to standard linear regression approach. However, this model cannot properly capture variation at the school level. The multilevel models with school weights are presented in columns (3) to (6). Using school weight only provides similar results to those obtained with student weights re-scaled with different methods. That should not be surprising for PISA-based research as sampling probabilities vary mainly by a school (depending, for example, on school size) and not within schools. On the other hand, the GK method, which simply assumes that weights within schools are equal to one, provides different results with a much lower estimate of school-level variance.

Based on this example and previous research in this area, one could conclude that a researcher should use survey weights at both levels and check if different scaling methods provide similar results or should apply school weights only (see Mang et al., 2021, for a similar analysis with analogous conclusions). Scaling to cluster size or effective cluster size should provide similar results for most circumstances. Ignoring weights or using other methods might be misleading, especially when a researcher is interested not only in fixed effects but mainly in variance decomposition and associations at different levels of cross-level effects.

The research also provides little guidance for models with more levels. It is also disputable how to apply survey weights when, for example, using a three-level model with countries added as an additional layer. In this case, a typical approach taken by OECD when analyzing PISA data is to re-scale student-level weights, to sum up to the same amount for each country. Thus, the final results could be interpreted as the OECD average, and country-level effects would explain how different policies affect this average. This is relatively straightforward to apply in a linear regression model, but as we saw in the above examples, re-scaling of student weights is not trivial in multilevel models and would affect estimates of variance components.

10.6 Estimation of Standard Errors

One of the reasons for using multilevel models when analyzing international student assessment data is that they recognize clustering of students within schools (or classrooms). The so-called robust standard errors that are adjusted for a correlation of student-level observations within education institutions are optional in some software packages for standard models like linear regression but are automatically applied in multilevel models. For many researchers, this is an advantage of multilevel approaches that they also cite as an argument for using such models with international student assessment data.

Studies like PISA or TIMSS, however, rely on complex sampling schemes and non-response adjustments. Unfortunately, detailed information on sampling schemes and survey weights adjustments is not available in the documentation for the reason of confidentiality. Participating countries often ask to hide key information in this respect from the public to make it impossible, for example, to estimate achievement for subnational entities or particular groups of students. Also, many countries' personal data protection law regulations disallow to provide detailed information on sampling when it might help identify individuals. Thus, dedicated solutions are applied in international student assessments to ensure that such requirements are met. For the same reason, variables used for complex sampling and response adjustments are not provided in the datasets, so it is impossible to correct survey weights or standard errors to reflect sampling design and non-response.

In practice, a difference between estimates of standard errors obtained from multilevel models and those obtained with a methodology developed by assessment organizers will vary by country and group of students analyzed. Thus, it is an empirical question, and estimates from linear regressions that fully follow the methodology developed by IEA or OECD experts can serve as a benchmark for multilevel models. In research that is based on simpler sampling designs, such discrepancies will usually be small. However, for complex surveys like PISA or TIMSS, they might be larger for countries with a lot of non-response, hidden stratification, or oversampling of some populations.

International large-scale assessments rely on resampling methods to estimate standard errors as these methods provide several advantages. The most important is that they can be used with many statistical models, even for which complex sampling data analytical solutions do not exist. In addition, replicate weights can incorporate confidential information about sampling and non-response without revealing any details to the public. Thus, in many surveys where privacy issues are at stake, this is a method preferred over providing sampling information in the datasets or in the documentation.

The replicate weights methods developed for educational studies mimic the sampling process by dropping individual schools (primary sampling units) in each replication. Thus, they provide standard errors that take into account sampling at the school level and clustering of student observations within schools. Multilevel models take that into account by directly modeling school-level effects. Combining two approaches makes little sense, and there is little research on this topic. In practice, however, replicate weights provide additional information in studies like PISA or TIMSS, which cannot be incorporated in the multilevel models. Thus, it is an important empirical question on how results from these models compare to those obtained with replicate weights methods. In practice, if both approaches provide different standard errors, then a researcher should analyze the sampling process and information incorporated in the survey and replicate weights more carefully. When standard errors estimated using replicate weights are larger, then caution should be taken when interpreting results from multilevel models as key information about sampling or non-response corrections might affect the results.

Table 10.5 provides a comparison of different methods for calculating standard errors for similar models. As before, the model explains student reading achievement using PISA 2018 data for all OECD countries. Columns (1) to (3) provide results for linear regression models, but with fixed effects for school, so the results can be compared with the multilevel model with random school effects. In column (1), standard errors are estimated as for simple random sampling, in column (2), standard errors are corrected for clustering at the school level (sandwich estimator), and in column (3), standard errors are estimated using the Balanced Repeated Replication method with Fay’s adjustment as advised in PISA technical reports (see OECD, 2020). These estimates of standard errors can be compared to those in column (4), which are estimated through the multilevel model with student and school weights and scaling to cluster size (the same model as in column 4 of Table 10.2).

Table 10.5 Comparison of standard errors obtained via different methods in linear regression and in a multilevel model—an example using PISA 2018 data for OECD countries

	(1)	(2)	(3)	(4)
	Linear regression	Linear regression with clustered standard errors	Linear regression with BRR standard errors	Student and school weight, scaling to cluster size
	b/se	b/se	b/se	b/se
ESCS index	15.5	15.5	15.5	16.8
	0.61	0.75	0.64	0.66
Constant	490.1	490.1	490.1	457.5
	0.49	0.20	1.11	2.31
School-level variance				3874.4
				199.04
Student-level variance				6180.8
				74.87
N of students	10,180	10,180	10,180	10,180
N of schools	249,334	249,334	249,334	249,334

Note All models are estimated with ten plausible values of reading achievement; dataset includes all OECD countries that have participated in PISA 2018 and have available reading achievement data

These results suggest that standard errors estimated with multilevel models are close to those obtained with the replicate weights method. In our example, more conservative are estimates with clustered standard errors in linear regression. However, a similar exercise should be performed in empirical applications to see if multilevel models provide standard errors that are more conservative than those obtained with replication weights and use additional, hidden information on the sampling process.

10.7 Multilevel Models with Additional Layers

Educational data have multiple layers, and depending on the research context and questions, these additional layers could be analyzed with multilevel models. The traditional choice of two-level models with students nested in classrooms or students nested in schools might not be optimal given that research questions might be related to other levels or interactions between these levels. For example, in research on education policy that uses international large-scale student assessment data, research questions often involve policies that are decided at the country level (e.g., the possibility of grade repetition or selection of students to different educational programs) but are applied at the school level, depending on individual teacher decisions, and are affecting relations between student-level variables and their achievement. For such research questions, it is natural to look at multilevel modeling as a perfect way to model these relationships. However, as we already saw, taking into account complex sampling is not straightforward even with two-level models. As countries vary in size and the number of schools sampled and in the population, a simple application ignoring survey weights could result in highly biased estimates.

It is questionable if country-level or any other level with a finite number of units could be modeled as a random effect. One could argue that countries or regions in which schools are nested represent observations from a superpopulation of all possible countries or regions (or policies possible to apply at these levels and randomly varying contexts). With obvious limitations of this approach, applications with more than two levels, including country or regional data, are interesting as they provide estimates of variance decomposition at these levels. For example, Grilli et al. (2016) estimate a four-level model to decompose achievement variance of 4th-grade students in Italy into the student, classroom, school, and province levels. The results show that achievement varies mostly at the individual level, and the province level is associated only with 5% or less of the overall variance. On the other hand, Hippe et al. (2018), using PISA data, show that achievement differences at the regional level in Spain and Italy are larger than differences in average achievement between EU countries. In a related paper, Hippe et al. (forthcoming) show that across the EU countries, the variance at the regional level is substantial when compared to the variance at the country level and that regional level predictors are strongly associated with regional level student achievement.

10.8 Summary

The data collected in large-scale international assessments are hierarchical in nature. The sampling scheme of these studies follows a general pattern of schools selected as primary sampling units followed by classrooms and students. Typical multilevel models applied to these data follow this sampling scheme with students nested in schools or classrooms. In this chapter, we discuss how to apply two-level models correctly with plausible values, survey weights at the school or classroom (or teacher) level, and scaling of survey weights at the student level. We also discuss issues related to standard error estimation when crucial information on sampling and non-response is hidden in replicate weights and not available for modeling in multilevel applications. Finally, we briefly discuss challenges in applying three-level models. In general, the methodology outlined in this chapter can be easily applied in popular statistical packages to properly analyze large-scale assessment data with two-level models. However, applying more complex multilevel models to these data still poses numerous challenges and requires caution when interpreting results.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (general), 149*(1), 1–26.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods, 35*(3), 439–460.

Avvisati, F. (2020). The measure of socioeconomic status in PISA: A review and some suggested improvements. *Large-Scale Assess Educ, 8*, 8. https://doi.org/10.1186/s40536-020-00086-x

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology, 9*, 49. https://doi.org/10.1186/1471-2288-9-49

Graubard, B. I., & Korn, E. L. (1996). Modeling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research, 5*, 263–281. https://doi.org/10.1177/096228029600500304

Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2016). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modeling of student achievement. *The Annals of Applied Statistics, 10*(4), 2405–2426.

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy &amp; Practice, 26*(4), 369–385.

Hippe, R., Jakubowski, M., & De Sousa Lobo Borges De Araujo, L. (2018). *Regional inequalities in PISA: The case of Italy and Spain*, EUR 28868 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-79-76296-3. https://doi.org/10.2760/495702, JRC109057.

Hippe, R., Jakubowski, M., & De Sousa Lobo Borges De Araujo, L. (forthcoming). *Regional variation of student performance in Europe: A multilevel model using unique PISA regional data.*

Jakubowski, M., & Gajderowicz, T. (2020). Student well-being factors: A multilevel analysis of PISA 2015 international data. *European Research Studies Journal, 23*(4), 1312–1333.

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115–147). Springer.

Lu, Y., & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-Scale Assessments in Education, 3*(1), 1–18.

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modeling: An investigation using PISA sampling structures. *Large-Scale Assessments in Education, 9*(1), 1–39.

OECD. (2009). *PISA data analysis manual: SPSS* (2nd ed.). PISA, OECD Publishing, Paris.

OECD. (2019). *PISA 2018 results (volume III): What school life means for students' lives*, PISA, OECD Publishing. https://doi.org/10.1787/acd78851-en

OECD. (2020). *PISA 2018 technical report*. OECD Publishing, Paris. Available at https://www.oecd.org/pisa/data/pisa2018technicalreport/

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B, 60*, 23–40. https://doi.org/10.1111/1467-9868.00106

Pitsia, V., Biggart, A., & Karakolidis, A. (2017). The role of students' self-beliefs, motivation and attitudes in predicting mathematics achievement: A multilevel analysis of the Programme for International Student Assessment data. *Learning and Individual Differences, 55*, 163–173.

Pokropek, A., Borgonovi, F., & Jakubowski, M. (2015). Socioeconomic disparities in academic achievement: A comparative analysis of mechanisms and pathways. *Learning and Individual Differences, 42*, 10–18. https://doi.org/10.1016/j.lindif.2015.07.011

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*, 805–827. https://doi.org/10.1111/j.1467-985X.2006.00426.x

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 1–17.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons Inc., New York.

Sun, L., Bradley, K. D., & Akers, K. (2012). A multilevel modeling approach to investigating factors impacting science achievement for secondary school students: PISA Hong Kong sample. *International Journal of Science Education, 34*(14), 2107–2125.

Sznitman, S. R., Reisel, L., & Romer, D. (2011). The neglected role of adolescent emotional wellbeing in national educational achievement: Bridging the gap between education and mental health policies. *Journal of Adolescent Health, 48*(2), 135–142.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record, 112*, 1008–1037.

Winnaar, L., Arends, F., & Beku, U. (2018). Reducing bullying in schools by focusing on school climate and school socioeconomic status. *South African Journal of Education, 38*(1).

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Yavuz, H. Ç., Demirtasli, R. N., Yalcin, S., & Dibek, M. İ. (2017). The effects of student and teacher level variables on TIMSS 2007 and 2011 mathematics achievement of Turkish students. *Egitim ve Bilim, 42*(189).

Maciej Jakubowski is a researcher in education and labor market policy and policy advisor. He holds a Ph.D. degree in economics and an M.A. in sociology from the University of Warsaw, where he works as an assistant professor. Between 2008 and 2012, he has worked for the PISA team at the OECD. Since 2012 served as an under-secretary of state at the Polish Ministry of Education. In 2014, he established Evidence Institute to promote evidence-based practice and support countries in analyzing international student assessments. His academic research focuses on large-scale student assessments and the methodology of policy evaluation.

Tomasz Gajderowicz is a researcher and policy advisor in the field of education and the labor market. Tomasz specializes in microeconometric methods for measuring incentives and preferences. Tomasz holds a Ph.D. in economic sciences and works as a consultant for the European Commission, World Bank, and other national and international institutions. He works as an assistant professor at the University of Warsaw and serves as a Research Director at Evidence Institute Foundation. He is the author of several publications about the transition from education to the labor market and research methodology.

Part IV
Multilevel Analysis of PISA and TIMSS Data

Chapter 11
Changing Trends in the Role of South African Math Teachers' Qualification for Student Achievement: Findings from TIMSS 2003, 2011, 2015

Heike Wendt, Daniel Kasper, and Caroline Long

Abstract Teacher education has an effect on the quality of instruction and consequently on students' learning, and achievement. Research in this area supports the presumption that the willingness to attend, for example, in-service training, is related to formal qualifications and work experience, but may also be affected by the current context. The focus of this paper is to understand these relationships in the South African context, where the enhancement of teacher qualifications has been identified as a core area of action; the policy on teacher education was renewed in 2007 and subsequently in 2014. The analysis is based on *Trends in International Mathematics and Science Study* data from 2003, 2011, and 2015. Multilevel regression models were calculated using SAS/STAT. Results show that South Africa has made substantial progress in uplifting teacher's formal qualification levels and in reducing structural inequality within its education system. Out-of-field teaching is still shown to be a common phenomenon and unevenly distributed. However, teachers with different qualification profiles do not differ in their usage of professional development opportunities. Also the teacher's formal level of education is in general not significantly associated with students' mathematics achievement. If at all, differential advantages are found for formal qualification rather than for specialization.

Keywords Professional development · Teacher qualification · Math teachers · TIMSS · Trends

H. Wendt (✉)
University of Graz, Graz, Austria
e-mail: heike.wendt@uni-graz.at

D. Kasper
University of Hamburg, Hamburg, Germany

C. Long
University of Johannesburg, Gauteng, South Africa

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_11

11.1 Introduction

The South African education systems under Apartheid showed high levels of racial segregation and inequality with regard to teacher education and teacher allocation to schools by qualification. The National Teacher Education Audit in South Africa revealed in 1995 high numbers of un- and under-qualified teachers as well as a highly fragmented provision of teacher education and training in a wide range of institutions, both in contact and distance mode (Hofmeyr & Hall, 1995). Many sources, such as the *President's Education Initiative* research project, have argued that the "most critical challenge for teacher education in South Africa has been the limited conceptual knowledge of many teachers". Consequently, the enhancement of teacher qualification has been a core area of action, and respective policy requirements were changed in subsequent years. According to the data presented in the DBE *Education for All* report (2014), 97% of all South African teachers in 2013, had the required teaching qualification, which was at that time a three-year post-school qualification, or REVQ1. However, the *National Policy Framework for Teacher Education* of 2006 uplifted the minimum requirement to a four-year degree or equivalent qualification. Arends (2013), using the representative TIMSS-Grade 9 data, showed that on average across the Grade-9-learners only 43% had completed a Bachelors' degree or equivalent in 2011, which was a substantial increase from 2002, where only 23% of all Grade-8-learners had been taught by teachers with at least a Bachelor degree or equivalent. According to Reddy et al. (2016) this percentage has significantly increased to 73% in 2015. South African and international research on professional knowledge (Adler et al., 2013; Carnoy et al., 2008; Darling-Hammond, 2000; Shulman, 1987) pointed out the importance of domain knowledge (subject matter knowledge, or content knowledge) and didactic knowledge (applied competence, knowledge-in-action-in-context) to support successful teaching practices. From this perspective when looking at teacher qualification, the subject matter knowledge obtained in a main subject area is additionally if not more important than a general qualification. In South Africa, as a consequence of teacher shortage, school characteristics, and a relatively high autonomy of schools with regard to staff appointments, it is not unlikely that teachers are placed in teaching positions that do not match their training, specialized qualifications, core knowledge, or skills. Qualitative research by Du Plessis et al. (2014) suggest that this phenomenon of out-of-field teaching is not an exception in South Africa and can have a strong influence on the quality of teaching and learning. Several studies conducted in the US and Australia as well as in Europe have researched the impact from the teachers' qualification on students' proficiency reflecting the view that teachers are a significant factor affecting children's learning outcomes (see Darling-Hammond, 2000). The majority of these studies have concluded that students who are taught by teachers with a subject-specific qualification achieve better results compared to those taught by out-of-field teachers (or those with a "lower" qualification), especially in mathematics and the sciences. For the South African context the commonness and distribution of out-of-field teachers has not yet been documented and, apart from few small-scale quality studies, has not been well researched. In addition, Adler et al. (2013) concluded in their review of research into policy and practice of teacher education in mathematics and science that there is a need for "empirical research that provides insight into the additional complexities of teaching both mathematics and the sciences across our diverse schooling conditions…" (p. 40). With this paper we aim to make use of data from a nationally representative monitoring study to describe trends in initial and continuing teacher education qualifications, making use of different criteria. In addition, we analyze the relationships between teacher qualification and student outcomes in diverse schooling contexts in 2002, 2011, and 2015.

11.2 Teacher Education of Mathematics Teachers in Post-apartheid South Africa

The South African education systems under Apartheid showed high levels of racial segregation and inequality. The National Teacher Education Audit in South Africa in 1995 found high numbers of un- and under-qualified teachers as well as a highly fragmented provision of teacher education and training in a wide variety of institutional sides, in both contact and distance mode (Hofmeyr & Hall, 1995). The Constitution of the Republic of South Africa, determined that all teacher education would henceforth fall under national control. Consequently, the teacher education institutional landscape since then has been completely transformed and concentrated in significantly fewer, larger, multidisciplinary public higher education institutions. However, the curriculum for teacher training programs was not centralized but rather fell within the responsibility of the faculties in which it was offered within the context of the expectations set up for teachers by the new schools' curriculum and the *Norms and Standards for Educators* (Sayed, 2004). Hence, the quality and content of the teacher training depended highly on the institution at which it was obtained.

Throughout the years both the curriculum as well as the *Norms and Standards* have been changed drastically (Reeves & Robinson, 2010). In 2000 the *Norms and Standards for Educators* regarded teachers who had obtained a three-year post-school qualification, or REVQ13,1 as adequately qualified. According to the revised *Minimum Requirements of Teacher Education Qualifications* policy (MRTEQ, DHET, 2015) of 2015, initial teacher education may follow two routes:

1. Complete a four-year Bachelor of Education degree. The requirement is to study two subjects. Mathematics Education as a study area in general comprises an academic component to strengthen the content knowledge of mathematics, and a professional didactics component to learn how to teach mathematics.
2. Complete an appropriate first degree in mathematics followed by a one year Advanced Diploma in Education, and register with the South African Council for Educators. Currently, the Advanced Certificate in Education is also used as professional development for mathematics teachers who have no teaching qualifications in this subject.

According to the *Integrated Strategic Planning Framework for Teacher Education and Development in South Africa 2011–2025* the relative shortage of teachers qualified and competent enough to teach mathematics is one of the major challenges to Teacher education for the upcoming years (DBE & DHET, 2011, p. 11). This field of action can be understood through the lens of education monitoring data. For example Arends (2013), using the representative TIMSS-Grade 9 data, showed that on average across the Grade-9-learners- only 43% had completed a Bachelors degree or equivalent, even though this is a substantial increase from 2002, where only 23% of all Grade-9-learners were taught by teachers with at least Bachelor degree or equivalent in mathematics. According to Reddy et al. (2016) this percentage has significantly increased since to 73%. However, the percentage of mathematics teachers, who have a Bachelor's degree in mathematics, has not yet been reported in detail.

In addition to the described Initial Professional Education of Teachers, the National Policy Framework for Teacher Education and Development identifies the Continuing Professional Development for Teachers (CPDT) as a complementary subsystem the purpose of which is to produce suitably qualified teachers. The core aim of CPDT for South African teachers is to enable learners to "learn well". This document "emphasizes the improvement of teachers' conceptual knowledge and skills through their PD". As part of a so-called Integrated Quality Management Systems (IQMS), a Developmental Appraisal System (DAS) was developed to guide and document the Continuing Professional Development of teachers (ELRC, 2003). The policy assumes that most teachers recognize the need for, and responsibility to, improve themselves professionally. But the IQMS includes evaluation through self-assessment and through the establishment of a development support group (DSG), which includes the teacher's immediate senior and a peer in their field of specialization. Teachers are allowed to choose which of their peers is to be part of the DSG (ELRC, 2003). All teachers have to earn PD points by attending accredited PD activities that meet their professional growth needs. However they can freely choose to obtain these through either formal or informal learning arrangements but are limited in their choice by available offerings. However the DAS has not yet has been formally implemented.

11.3 Teacher Allocation and Placement in Schools

Under the apartheid government, spending on school education varied greatly depending on the race categorization of the school. This resulted in an imbalance in terms of facilities and equipment but also in terms of staffing and teaching posts available at the school (Jansen & Taylor, 2003). The new ANC government issued a series of Acts that prohibited the use of unfair discrimination on the one hand and facilitated the promotion of teacher integration on the other, aiming to open the door of equal opportunity and access to all teachers, irrespective of race or gender, to apply for any position at any school provided that they satisfied the requisite employment criteria (Mampane, 2009). Among the series of Acts, the South African Schools Act (SASA), Act 84 of 1996 guaranties a high degree of decentralized governance authorities to individual schools, in the form of School Governing Bodies, including playing a major role in the selection and appointment of teachers. According to the SASA the governing bodies of all public schools have the right to make recommendations to the relevant (provincial) Head of Department regarding the appointment of educators to the schools for whose governance they are responsible. In addition, school governing bodies may also "establish" posts for educators. In short schools in consultation with the (provincial) Head of Department have the right to recruit and appoint teachers.

However, Onwu and Sehole (2010) argue that as a consequence of a general teacher shortage, and the differing attraction of work environments, rural communities are disadvantaged in terms of recruiting and retaining qualified teachers. The Teacher Rural Incentive Scheme (TRIS) put in place in 2008 to address this issue is argued and found by Poti et al. (2014) to not sufficiently address this inequality of provision. Mampane (2009) in addition found in her qualitative study that different stakeholders understand and interpret legislation on teacher appointment differently because of their own experience, training, academic qualification, and socio-historical factors and may therefore make different appointment and allocation decisions. However as Reeves and Robinson (2010) argue, both the lack of qualified teachers as well as the fact that many in-service teachers were professionally trained under valid criteria of the time, it is in practice additionally difficult to identify whether the formal teachers' professional qualifications and the subject specializations in their diplomas or degrees qualify them to teach a specific learning area or subject and phase level. Difficulties in matching teachers to posts in schools may in addition arise out of timetabling issues, and/or limitations in staffing allocations, particularly in small schools with low staff ratios.

Hence, school characteristics and a relatively high autonomy of schools with regard to staff appointments make it not unlikely that teachers are placed in teaching positions that do not match their training, specialized qualifications, core knowledge, or skills. They may be allocated some teaching responsibilities that are only partially within their field of expertise and be required to teach areas or subjects at levels out of their field of expertise.

11.4 Research on Teacher Participation in Professional Development Programs

Professional development of teachers can be understood as a life-long learning process comprising "formal and informal support and activities that are designed to help teachers develop as professionals. This includes taught courses and in-school training, as well as activities such as coaching, mentoring, self-study and action research" Coldwell (2017, p. 189). Many in-service education/teacher professional development programs are delivered in the form of workshops, seminars, conferences, or courses. They have been criticized by many researchers in developing and developed countries as being brief, fragmented, incoherent encounters that are decontextualized and isolated from real classroom situations (Steyn, 2011). Westheimer (2008) argues that teachers are capable of compiling relevant learning material themselves and do not require external assistance, whereas Morris et al. (2003) consider a combination of both external and internal professional development to be most effective. Focusing on out-of-field teachers Hobbs (2013) argues that professional development programs constitute one central support mechanism for out-of-field teachers.

Reddy et al. (2015) report on the basis of data from the TIMSS study 2011 that South African teachers, compared with an international average, have attended a high number of professional development activities especially in the areas of mathematics curriculum, mathematics content, critical thinking and assessment (p. 29). Mullis et al. (2016) report similar results for TIMSS 2015: Four out of five South African Grade-9-learners are taught by a teacher that reported to have participated in professional development in mathematics curriculum (86%) or content (84%) over the past two years, followed by 73% for mathematics assessment, 58% for mathematics pedagogy, 56% critical thinking and problem solving, 52% addressing individual students needs and 45% integrating modern technologies. Trends of these data have not yet been comprehensively reported.

Several factors are supposed to be considered, when looking at the participation in professional development programs. Teacher career stage models, such as the prominent model by Huberman (1989) that differentiates between five stages of teachers' development, point out that teacher orientations, engagement, and professional expertise as well as the need and willingness of participation in activities may develop through their career, as the different stages are associated with different challenges. For South Africa the Educator Workload Study of 2005 found that on average in 2005 teachers spent 2.24 h a week on professional development activities (5.4% of their total time was spent on school-related work and teaching. The study did specifically look into differences in different patterns of participation in PD activities, but found overall that for the time spent on educational activities there were significant differences by subject, age-group, and sex (Chisholm et al., 2005). Research also suggests that teachers in addition may also have different preferences (Robinson & Carrington, 2002) and interests. Here two different mechanisms for teacher participation in professional development programs can be distinguished. First teachers may follow their interests and choose training in a subject that corresponds to their major to expand their professional competencies in that field (e.g., Desimone et al., 2006) Alternatively, based on the identification of personal deficits, teachers may feel the need to attend training as they provide structured learning opportunities in order to obtain necessary additional subject knowledge ("deficit hypothesis"). In addition Steyn (2011) argues that guidance and support by the school leadership, collaboration structures in the school contexts, and school culture may play a crucial role in decisions on how teachers may choose to augment their professional development.

11.5 Research on the Relationship Between Teacher Qualification and Learning Outcomes

Several studies conducted in the US and Australia as well as in Europe have researched the impact of the teachers' qualification on students' proficiency (e.g., Darling-Hammond, 2000; Dee & Cohodes, 2008; Goldhaber & Brewer, 2000; Monk & King, 1994; Richter et al., 2012, 2013), reflecting the view that teachers are a significant factor affecting children's learning outcomes. The majority of these studies have concluded that students who are taught by teachers with a subject-specific qualification achieve better results compared to those taught by out-of-field teachers (or those with a "lower" qualification), especially in mathematics and the sciences. However, one needs to mention that "most of the studies that find statistically significant relationships between teacher training and student achievement find that the effects of these characteristics are small and specific to certain contexts" (Goldhaber, 2002, p. 54). In addition, the studies use different indicators of the teachers' formal qualification, which can firstly be explained by the diverse and ongoing changing educational systems even within each country and, secondly, by the availability of the data that may have been provided, for example, by a governmental institution for purposes different from evaluating the relationship between teacher qualification and student performance. Thus, the comparability of these studies is limited and the context of each study needs to be carefully examined.

As Porsch and Wendt argue the methodology of the existing quantitative studies on the question of whether the formal qualification of teachers can explain variance in the students' proficiency can be described in three ways: (1) A comparison between mean scores is presented as differing between two or more groups of students that are based on the different qualifications by their teachers (e.g., May, 2006; Richter et al., 2012), (2) regression analysis has been conducted without modeling the different levels but exclusively considering teacher variables (e.g., Goldhaber & Brewer, 1996). Others have also controlled for student characteristics; these latter studies either include sole indicators of the teachers' qualification (e.g., Dee & Cohodes, 2008) or, in addition, further teacher characteristics like experience (e.g., Monk & King, 1994), gender, or ethnic background (e.g., Clotfelter et al., 2006, 2012). (3), A further approach is the application of multilevel regression modeling whereby teacher and student characteristics are included (e.g., Richter et al., 2013). In addition, some researchers provide interactions between student and teacher characteristics (e.g., Zuzovsky, 2009). This type of regression analysis is recommended for nested data if the number of participants on each level is sufficient. This method would lead to a more accurate picture of the results since the standard errors of the parameters would not be underestimated and there would be separation of the variance attributed to the levels (e.g., Hox, 2010).

For developing countries, the relationship between teacher qualification and learning outcomes is not very well researched. Studies focus rather on the general role of teacher education levels and other related teacher characteristics. Glewwe et al. (2011) found in their meta-analysis examination of 43 high quality studies, very ambiguous results when looking at teacher education levels. They concluded that there is little evidence that teachers' level of education has any impact on student test scores; however, some evidence was found that teacher experience has a moderate, and teachers' education and experience a strong positive effect, when measured more directly as knowledge of the subjects that they teach. In-service teacher training was also found to have a positive impact on students' test scores. However, the only two South African studies by Gustafsson (2007) and Van der Berg (2008) that are included in the meta-analysis did not explicitly look at teacher education factors, therefore on the basis of this meta-study it remains unclear to what extent results can be generalized for the South African context.

11.6 Research the Relationship Between Teacher Qualification and Learning Outcomes in Mathematics in South Africa

In South Africa the importance of teacher qualification has been stressed for many years (Taylor & Vinjevold, 1999) highlighting research showing teachers' low levels of conceptual and content knowledge of their subjects. Although the importance of the issue may be relegated as common sense, there is as yet little conclusive research evidence as to what kind of measurable effect teachers' knowledge has on learning (if any) in this country even though the enhancement of teacher qualification has been a core area of action since the democratic government was elected in 1994.

A study undertaken in South Africa in 2001 by Crouch and Mabogoane, found that teacher qualifications as a measure of teacher quality were strongly associated with an increase in learner pass rates in the school-leaving examination. It should be noted that the study was conducted at a time when difference between qualified and un/under-qualified teachers was stark, as some 36% of the teachers in South Africa were unqualified (had no professional teaching qualification) or under-qualified (had less than a three-year tertiary qualification). In 2005 the Educator Workload Study (Chisholm et al., 2005) showed a positive relationship between qualification level and time spent on teaching-related activities. A study done by Kunene (2013) based on TIMSS 2003 data, which did not explicitly consider teacher qualification but the relationship of other related teacher characteristics and mathematics achievement found a significant relationship with teacher experience. However, it should be noted that from today's perspective, all three studies have some methodological limitations.

More recent large-scale education data do not provide clear results but hints at the presumption that teacher qualification is positively related to student achievement: This is also the case in South Africa. The descriptive bivariate statistics reported in the TIMSS International Report of 2015 (Mullis et al., 2016), comparing student achievement for groups of learners taught by teachers with different majors, shows up to 30 score point differences in favor of groups taught by teachers that majored in mathematics education. However, these differences are associated with large standard errors and are not significant.

A small quasi-experimental study conducted by Carnoy et al. showed that teacher quality and opportunity to learn were estimated to have significant positive effects on learner gains in mathematics test scores even though the effect size was very small. In addition, two secondary analyses of the South African Grade-6-SACMEQ data that studied the relationship of teacher knowledge and student mathematics achievement using advanced modeling techniques found no general effect but rather that teacher knowledge only had a significant positive relationship with learner's performance in the wealthiest quintile of schools (Shepherd, 2013; Spaull, 2011), that is schools that are likely to be independent or that fell under the former white and Indian education departments.

Given the stark variety of learning contexts in South Africa it is plausible to assume that the task of creating effective learning situations in the classroom based on the individual learning requirements previously described as adaptive teaching competence (Beck et al, 2008), is a critical part of teachers work. Since classes differ in their composition, teachers need to adapt to the different working conditions and are not equally able to make use of their subject-specific training. In fact, Monk and King (1994) on the basis of their study using data on 2,829 students from the Longitudinal Study of American Youth suggest with their findings that the qualification in one teaching subject may not affect low- and high-performing students in the same way. Thus, one can assume that students with different educational needs may profit differently from the qualification of their teachers, which has also been revealed in classroom observations of out-of-field teachers (cf. du Plessis, 2013). A study recently published by Arends et al. (2017) found substantial differences in classroom practices between fee-paying and no-fee paying schools using multilevel analysis on TIMSS 2011 data for South Africa. This study also revealed a positive association between teachers' high endorsement of the selected classroom practices and learner performance. A video study of classroom practices of mathematics teachers in 40 schools in Gauteng, the most highly urbanized of South African provinces suggests that it is plausible to assume an interaction between composition factors and teacher qualification. The authors argue that their study provided some evidence "that (the institution) where teachers took their pre-service training (most were trained before 1994 in teacher training colleges that have since been closed) may have an impact in how much mathematical pedagogical content knowledge they have, and that their mathematical pedagogical content knowledge is related positively to the quality of their mathematics teaching" (Carnoy et al., 2008).

11.7 Research Questions

1. What is the percentage of teachers with different qualification levels teaching 8/9th Graders in South Africa using the TIMSS datasets. Has this percentage changed since 2002?

2. When considering school context factors, are there different patterns of teacher allocations with regard to their qualification profile? Has this pattern changed since 2002?
3. To what extent do South African mathematics teachers participate in formal professional development opportunities? Are there different patterns for different qualification profiles? Have these patterns changed since 2003?
4. Are there significant differences between students' test scores depending on the qualification of their teachers? Have these relationships changed since 2003?

Hypotheses

1. Arends (2013), using the TIMSS data, showed that across the ninth grade only 43% had completed a Bachelors degree or equivalent, which is a substantial increase since 2002, where only 23% of all Grade-8-learners were taught by teachers with at least Bachelor degree or equivalent. According to Reddy et al. (2016) this percentage has significantly increased since 2013 to 73%. Building on Du Plessis et al. (2014) we expect to find a substantial number of teachers that are out-of-field, meaning they taught mathematics to the TIMSS class of that year, but had not specialized in mathematics or mathematics education throughout their initial teacher training. In accordance with a general positive pattern, we do expect a decrease of that percentage over time. Combining these two factors, we identify teachers which can under the current legislation be considered as fully qualified mathematics teachers,[1] and we expect corresponding findings.
2. Du Plessis et al. (2014), Mampane (2009), and Onwu and Sehole (2010) argue that the placement of teachers in positions is dependent on school context factors, whereas there is no systematic allocation policy based on qualification criteria. This mostly qualitative research showed that schools that are more privileged might have more opportunities to choose between teachers. We therefore expect to find moderate differences in the number of fully qualified teachers among the more privileged schools. We also expect a lower percentage of fully qualified teachers in rural schools.
3. As the policy on teacher education was renewed in both 2007 and 2014, with a resulting increase in the demands on teacher qualification levels, we expect to find a general increase in development activities. As especially teachers with a qualification degree below Bachelor level are explicitly targeted to upgrade their degree in formal teacher education programs we expect them to a have a higher involvement in such activities (deficit hypothesis). In accordance with the interest hypothesis we expect to find a higher involvement especially for those teachers who obtain a qualification in mathematics throughout their initial teacher training. Since the curriculum has been changed several times since 2002

[1] Min. Honors/Bachelor degree and mathematics (education) as a majors study area.

we expect a high and increasing participation rate of mathematics teachers in training that focus on curriculum and content. Since the Revised National Curriculum Statements between 2002 and 2011 demanded that teachers follow a new pedagogical approach to mathematics teaching, we additionally expect to find an increase in participation activities in this area. As data suggests an increase in the availability of new information technology in schools we expect to find an increasing number of teachers participating in PD-courses around questions of how to facilitate an integration into mathematics teaching.

4. The relationship of teacher qualifications and students' test scores in mathematics is not well researched. Building on the work of Crouch and Mabogoane we expect to find a relationship between teacher qualification and students achievement in mathematics in TIMSS 2003 to the disadvantage of students taught by teachers with a lower formal qualification level. For the more recent studies in accordance with Mullis et al. (2016) we expect findings to be similar but perhaps as the results from Shepherd (2013) and Spaull (2011) suggest teacher qualification may not have an overall effect but rather play a role depending on the school context.

11.8 Data, Population, and Sample

We use the International Association for the Evaluation of Educational Achievement's (IEA) cross-sectional Trends in Mathematics and Science Study (TIMSS) datasets for the years 2003, 2011, and 2015. These comparative large-scale studies aim to monitor and evaluate educational performance internationally, and to develop an evidence base for stakeholders of the factors that are associated with more equitable educational equality. Therefore, in addition to the assessments in mathematics and science, extensive questionnaires are administered to principals, teachers, and students. Our target population is mathematics teachers of Grade 8/9 learners in South Africa, which is captured by implementing a two-stage stratified sampling design. The result is a nationally representative sample for the school and student population. In Table 11.1, we show the sample sizes for the student, school, and mathematics teachers per study circle.

Table 11.1 Sample characteristics

	TIMSS 2003	TIMSS 2011	TIMSS 2015
Target group	Grade 8	Grade 9	Grade 9
N_{schools}	255	285	292
N_{teachers}	256	318	327
N_{students}	8952	11,969	13,708

11.9 Measures

We only included measures in our analysis which were administered with equivalent phrasing in all three TIMSS study cycles. As we used the dataset available as part of the international database we were limited in our indicator selection as national additions to the questionnaire, which might have been used in other studies, are not publicly available.

11.9.1 Teacher Qualification

We distinguish four groups: (1) Teachers with a university degree that studied mathematics or mathematics education as a major. Teachers with this qualification profile can be under the current legislation be regarded as fully qualified. (2) Teachers with a teaching qualification in mathematics obtained at an institution below university level. Teachers with this qualification profile are specialized mathematics teachers and were under previous legislation also considered to be fully qualified. (3) Teachers with a university qualification, that majored in a subject other than mathematics. (4) Teachers that have a qualification below university level, that majored in a subject other than mathematics. In line with the definition by Ingersoll according to which "out of field educators are sufficiently trained educators, well qualified but placed in teaching positions that do not match their training, specialized qualifications, core knowledge, skills, beliefs, values or approach", we consider teachers with qualification profiles 3 and 4 as out-of-field teachers. The indicators were created combining answers from teacher self-reports in the TIMSS teacher questionnaire. Here teachers were asked: "*What is the highest level of formal education you have completed?*" and "*During your <post-secondary> education, what was your major or main area(s) of study?*". Teachers that reported having a degree on or above "Bachelor's or equivalent level—ISCED Level 6> " and studies "Mathematics" or "Mathematics Education" were judged to belong to group 1. Teachers with a degree below the respective level but with a major in "Mathematics" or "Mathematics Education" to Group 2, and so forth.

11.9.2 Teacher Covariates

Building on previous research, we decided to include experience, the participation in professional development activities as well as the sex of the teachers. Even though Glewwe et al. (2011) found only weak evidence that teacher experience had a positive impact on pupils' test scores, we decided in accordance with research done on teacher career stage to include answers on the experience ("By the end of this school year, how many years will you have been teaching altogether?") not as a metric variable but to distinguish for different age groups: 0–5 years, 6–10 years, 11–20 years and more than 20 years.

As indicators for professional development we use answers to a question asking specifically in which of six given areas a teachers has participated in professional development activities over the last two years (*In the past two years, have you participated in professional development in any of the following?*). It can be assumed that formal activities are meant but it is not clearly stated. For research question 3, we use the single items separately, whereas for the regression analysis we simply added the number of areas positively checked. Not knowing anything about the intensity, depth, or the quality of the developmental activity, we assume teachers who engaged with relatively more topics in their professional development have been relatively more engaged than colleagues who did not.

11.9.3 Student Level Control Variables

A number of studies have already investigated the relationship between student and home factors with mathematics achievement in the South African TIMSS datasets of 2002, 2011, and 2015 (Reddy et al., 2012; Visser et al., 2015). Building on this work we decided to include sex, self-concept, language, parental education, home educational resources (own study desk and number of books) as well as the availability of electricity and water, as socio-economic home indicators.

11.9.4 Classroom Context Control Variables

Studies have also shown that it is important to include school factors when investigating the relationship of teacher factors and mathematics achievement (Shepherd, 2013; Spaul, 2011). The work of Winnaar et al. (2015) and Prinsloo and Rogers (2013), have shown that both socioeconomic as well as language should be taken into account. Building on this work we decided to include three indicators class size, proportion of students that do not speak the test language at home and the proportion of students from families living in poverty. We therefore calculated a “poverty index” assuming that students reporting that they do not have electricity, running tap water, more than 25 books, a study desk nor an own room, are more likely to live under poverty conditions than students that report to have them.

11.9.5 Student Outcomes

Our outcome measures are mathematics (“MAT”) that is the mathematics achievement of the Grade 8/9 students on the overall mathematics scale.

In Table 11.2, we show each of the measures we use for the analyses including their scales, measures of central tendency and spread, minimum and maximum values, and the theoretical constructs operationalized.

11.9.6 Analytic Strategy

For research questions 1 to 3 we calculated descriptive statistics and linear regression using the IEA IDB Analyzer, which takes into account the complex sampling design using sampling weights and replicate weights. To proceed to answering research question 4, we approach student achievement and school effectiveness within the contextualized achievement model. In so doing, we fit four two-level linear hierarchical models to the cross-sectional data taking into account a broad set of student background and classroom characteristics, and the teacher qualification measures and teacher covariates. The models were identical across study cycles so that changes between coefficients can be interpreted. A null model serves as a reference for the interpretation of the variances explained by the independent variables. In Model 1 we regress the teacher qualification profiles on student achievement. Here as in the following models students taught by fully qualified mathematics teachers serve as a reference group. In Model 2 we introduce student level predictors as well as class composition variables, as policy, previous research and the finding presented in Table 11.2 indicated that teachers are not allocated to schools randomly. In Model 3 we additionally introduce teacher sex, experience, and participation in professional development as covariates as research suggest that these might be relevant covariates. In Model 4 we additionally introduce an interaction term for the teacher qualification profiles and a composition indicator to study possible differential effectiveness which both qualitative and quantitative studies suggested. Multilevel modeling enables us to account for the nested nature of students within classrooms and the classroom within schools, which, if left uncorrected for, could lead to biased estimates (Murnane & Willet, 2011). These multilevel regression models are fitted to the data using %SURVEYHLM, a SAS®[2] software macro for multilevel analysis with large-scale educational assessment data (Kasper et al., 2018). Also for these calculations the complex sampling and assessment design is acknowledged and sampling and replicate weights, as well as the five plausible values used. Listwise deletion was the strategy used to deal with missing data. For all of our models, we set our significance level at $\alpha = 0.05$, accepting the convention that our null hypotheses will be rejected 5% of the time when they are, in fact, true. As only relationships with covariates are explored and not overall achievement trends, we assume the sampling differences with regard to the Grade level (8th vs. 9th Grade) do not influence the results.

[2] *SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.*

Table 11.2 Descriptive statistics of indicators used (percentages of Grade 8/9 learners in percent)

			TIMSS 2003				TIMSS 2011				TIMSS 2015			
Class Level														
Teacher Characteristics			%	(SE)			%	(SE)			%	(SE)		
University Ed & Major (Mathematics)			23,6	(3,6)			51,4	(3,9)			62,2	(3,2)		
Non-University Ed & Major (Mathematics)			55,0	(4,0)			36,5	(3,9)			23,8	(2,7)		
University Education: Other Spec.			13,5	(2,1)			8,5	(2,2)			10,0	(2,2)		
Non-University Ed & Other Spec.			7,9	(2,1)			3,7	(1,1)			3,9	(1,1)		
			100,0				100,0				100,0			
Experience; 0-5 y			20,3	(2,8)			22,4	(3,2)			25,4	(3,4)		
Experience; 6-10 y			33,7	(3,8)			18,5	(3,1)			21,1	(3,1)		
Experience: 11-20 y.			36,6	(3,8)			37,1	(3,4)			27,2	(3,4)		
Experience: at least 20 y.			9,4	(2,5)			22,0	(3,3)			26,3	(3,4)		
			100,0				100,0				100,0			
Sex: female			39,5	(3,4)			42,2	(3,8)			46,2	(3,7)		
	Min.	Max.	*M*	(SE)	*SD*	(SE)	*M*	(SE)	*SD*	(SE)	*M*	(SE)	*SD*	(SE)
Intensive participation in PD	0	6	3,4	(0,18)	2,1	(0,07)	3,5	(0,17)	2,0	(0,08)	4,0	(0,12)	1,8	(0,08)
Class composition														
Class size [(a)]	4	99	35,6	(0,49)	8,0	(0,93)	42,5	(0,90)	15,1	(1,25)	42,8	(1,22)	15,0	(0,89)
Small proportion of students from families living in poverty[(as)]	9	100	54,0	(1,03)	15,6	(0,90)	60,1	(0,66)	14,0	(0,76)	62,6	(0,74)	12,4	(0,64)
High proportion of students speaking the test language at home [(as)]	0	100	27,5	(1,86)	30,3	(1,58)	26,0	(0,99)	28,2	(0,90)	30,8	(1,58)	28,4	(0,82)
Student level														
Age	9,7	19,9	15,1	(0,04)	1,3	(0,02)	16,0	(0,04)	1,2	(0,02)	15,7	(0,03)	1,2	(0,02)
High self-concept	0	3	2,1	(0,02)	0,7	(0,01)	2,2	(0,01)	0,7	(0,01)	2,1	(0,02)	0,7	(0,01)
			%	(SE)			%	(SE)			%	(SE)		
Sex: female			51,4	(0,9)			48,0	(1,0)			51,3	(1,1)		
Speaks (almost) always test language at home			27,7	(1,9)			26,1	(1,0)			30,9	(1,6)		
Parents highest Education level max. Lower Sec. Education			45,5	(1,3)			28,7	(1,1)			20,6	(1,0)		
Parents with max. upper second. Education			43,2	(0,9)			52,4	(0,9)			53,0	(1,1)		
Parents with min. University degree			11,3	(1,0)			18,9	(0,7)			26,4	(1,6)		
			100,0				100,0				100,0			
Own study desk at home			57,8	(1,5)			56,2	(0,8)			61,3	(1,0)		
More than 25 books at home			24,8	(1,3)			22,6	(0,6)			20,6	(0,9)		
Electricity at home			75,8	(1,8)			86,6	(1,4)			90,7	(1,1)		
Running tap water at home			60,2	(1,5)			71,7	(1,4)			73,9	(1,2)		

(a) aggregated nr. of students; (s) student reports; (t) teacher reports; (as) aggregated student responses

11.9.7 Limitations

For our study of the relationships we use cross-sectional data, so no causal interpretations of found relationships are possible. In addition our research design uses indicators obtained from the teacher questionnaire administered in the TIMSS study. It should be noted that TIMSS is not a representative teacher survey, as students serve at the target group. Given this sampling the data only allows interpretations on the student level. Given the cross-sectional design of the study, the interpretation is limited by the fact that nothing is known about the education history of students and as a result it may well be that students for a substantial time throughout their education careers may have been taught by teachers with different qualification profiles. In addition our analysis limits distinguishing the "effects" of teacher qualification. For example, we do not consider the role that other structural factors (such as school location), school environment, school management and teacher instruction play in achievement or in moderating, or even mediating an "effect" of teacher qualification. A further limitation due to omitted variables bias comes from our exclusion of explicit measures of teacher qualification, personality traits, and acquired knowledge in other formal or informal learning opportunities.

11.10 Results

11.10.1 Percentage of Learners by Qualification Levels

In Table 11.3, we show the changes in qualification levels of mathematics teachers between 2002, 2011, and 2015 (percentages of Grade 8/9 learners in percent) using different indicators for qualification. A significant increase of learners being taught by teachers that studied mathematics, teachers with a university degree since 2002 becomes apparent. The number of students taught by fully qualified teachers in mathematics increased from about 24% in 2002 to about 62% in 2015. The number of students taught by out-of-field teacher also decreased significantly. However in 2015 still about 14% of all Grade 9 students were taught by teachers in mathematics that did not specialize in mathematics throughout their initial teacher training.

11.10.2 Teacher Allocations by Qualification Profile and School Student Composition

Our second research question was concerned with the distribution of teachers by school context. In Table 11.4, we show the distribution of fully qualified mathematics teachers by school locations. This indicator was only administered in 2011 and 2015 so results can only be shown for these years. Table 11.4 shows nominal differences

Table 11.3 Changes in qualification levels of foundation phase mathematics teachers between 2002, 2011, and 2015 (percentages of Grade 8/9 learners in percent)

	TIMSS 2003		TIMSS 2011		TIMSS 2015		Changes	
	$\%_{02}$	(SE)	$\%_{11}$	(SE)	$\%_{15}$	(SE)	$\%_{15}–\%_{02}$	(SE)
Mathematics or mathematics education was a main study area	68,7	(3,2)	82,5	(2,5)	85,9	(2,3)	17,2*	(3,9)
Min. Bachelor/Honors degree	37,2	(3,5)	59,9	(3,9)	72,3	(2,9)	35,1*	(2,9)
University Ed & Major (Mathematics)	23,6	(3,6)	51,4	(3,9)	62,2	(3,2)	38,6*	(4,8)
University Education: Other Spec	13,5	(2,1)	8,5	(2,2)	10,0	(2,2)	-3,5	(3,1)
Non-University Ed & Major (Mathematics)	55,0	(4,0)	36,5	(3,9)	23,8	(2,7)	-31,2*	(4,8)
Non-University Ed & Other Spec	7,9	(2,1)	3,7	(1,1)	3,9	(1,1)	-3,9	(2,4)

* = significant change ($p < 0.05$)

Table 11.4 Distribution of fully qualified mathematics teachers by school locations in 2011 and 2015 (percentages of Grade 8/9 learners in percent)

School location reported by principals	TIMSS 2011			TIMSS 2015		
	$\%_{\text{Schools}}$	$\%_{\text{Expert}}$	$(\text{SE}_{\text{Expert}})$	$\%_{\text{Schools}}$	$\%_{\text{Expert}}$	$(\text{SE}_{\text{Expert}})$
Urban, densely populated	14,0	47,5	(9,4)	14,2	**78,4**	(7,6)
Suburban, on fringe of urban area	12,2	**76,2**	(7,9)	15,1	66,4	(6,9)
Medium size city or large town	12,4	50,7	(8,1)	8,2	54,8	(11,0)
Small town or village	27,0	38,1	(7,2)	28,7	57,1	(6,8)
Remote Rural	34,4	36,9	(7,4)	33,8	58,9	(5,6)

Bold = significantly higher than in all other groups ($p < 0.05$), TIMSS 2015 except “Suburban”

between the percentages of learners taught by fully qualified mathematics teachers by school location. For 2011 it can be seen that learners schooled in suburban locations were significantly more often taught by fully qualified teachers than their peers going to school in a different location. For 2015 it is found that learners schooled in urban locations were significantly more often taught by fully qualified teachers than their peers going to school in medium size cities, towns, or remote rural location. Against

Table 11.5 Distribution of fully qualified mathematics teachers by quintiles "relative poverty" in 2002, 2011, and 2015 (percentages of Grade 8/9 learners in percent)

Class composition Quintile	TIMSS 2003		TIMSS 2011		TIMSS 2015	
20% of classes with	%$_{02}$	(SE)	%$_{11}$	(SE)	%$_{15}$	(SE)
1 many disad. students	16,4	(6,0)	43,3	(8,7)	57,2	(7,2)
2	27,9	(8,4)	38,7	(8,5)	60,2	(7,7)
3	16,6	(6,4)	47,6	(8,2)	65,2	(7,3)
4	12,2	(6,2)	50,8	(7,6)	52,8	(8,3)
5 many privil. students	29,5	(7,0)	54,4	(5,8)	73,5	(6,0)

All differences not statistically significant ($p < 0.05$)

our expectation, we do not find a systematic disadvantage of schools in remote rural areas.

In Table 11.5, we show the distribution of fully qualified mathematics teachers by schools grouped into quintiles according to their school composition. Nominal differences in favor of schools with a higher number of students from affluent backgrounds become apparent, but they are not significant.

11.10.3 Intensity and Focus of Teacher Participation in Formal Professional Development Activities

Our third research question was concerned with participation in formal professional development opportunities. In Table 11.6 we show the average number of mathematical topics covered in professional development programs by teachers in 2002, 2011, and 2015. In 2015 on average mathematics teachers in South Africa engaged with four different topics of mathematics education over a period of two years. This engagement is a significant increase since both 2002 and 2011.

In a next step we used linear regression analysis to test if teachers with different qualifications profiles show a different degree of involvement in their professional development activities. We calculated separate models for 2002, 2011, and 2015 and

Table 11.6 Participation in professional development in mathematics over two years in 2002, 2011, and 2015 (Average number as reported by teachers, six topics given covered)

TIMSS 2003		TIMSS 2011		TIMSS 2015		Changes			
M_{02}	(SE)	M_{11}	(SE)	M_{15}	(SE)	M_{15}–M_{02}	(SE)	M_{15}–M_{11}	(SE)
3,4	(0,2)	3,5	(0,2)	4,0	(0,1)	0,7*	(0,2)	0,6*	(0,2)

* = significant change ($p < 0.05$)

did not find any significant differences between the teachers with different qualification profiles, and also no significant differences when we controlled for teacher experience and sex.

In Table 11.7 we show the PD-participation profiles of mathematics teachers over two years in each of 2002, 2011, and 2015 by area. Replicating the findings already published by Mullis et al. (2016) and Reddy et al. (2015) it can be seen that mathematics content, mathematics curriculum and assessment are the topic areas more commonly chosen by mathematics teachers rather than pedagogy, critical thinking/problem solving or the integration of information technologies. However it should be noted that even less frequently chosen topic areas are according to the self-reports still covered by at least a third if not half of the teachers, depending on the study cycle. Looking at similarities over time we find that Content, Curriculum, and Assessment are the most frequently covered areas, but with different emphasis: In 2002 Assessment was among the three areas more often covered, whereas in 2011 no notable differences were found, and in 2015 Curriculum and Content were more frequently chosen. Looking explicitly at changes over time we find a significant increase in the usage of professional development opportunities for four areas, especially mathematics curriculum and mathematics content and to a minor extent in pedagogy and integration of information technologies.

In a next step, see Table 11.7, we differentiated by teacher qualification profiles to learn if teachers depending on their qualification profiles have different interests or needs in terms of professional development and may therefore show differences in relative relevance of the topics. In Table 11.8 we show the PD-participation profiles over two years in each of 2002, 2011, and 2015 by area and qualification profiles. Overall, we find great similarities across all four groups of teachers: Over all study cycles for all teachers, regardless of their qualifications, Curriculum, Content, and Assessment are the topic areas most frequently covered.

11.10.4 Relationship Between Teacher Qualification and Student Outcomes

Our fourth research question was concerned with the relationship between teacher qualification and student achievement and possible changes in these over time. In Table 11.9 we present the results of the multilevel analysis with student's mathematics achievement as the dependent variable.

Based on the results for Model 2 and 3 where teacher qualification profiles are regressed on student mathematics achievement, controlling for differences in classroom composition, student and teacher characteristics, we cannot reject the null hypothesis and conclude that on average in South Africa differences in formal qualification profiles do not have statistically significant, predictive relationships with mathematics achievement of Grade 8/9, net of all else in all three study cycles. An exception is TIMSS 2003 where students taught by out-of-field teachers without a

Table 11.7 Participation in professional development in mathematics over two years in 2002, 2011, and 2015 by topic (percentages of Grade 8/9 learners in percent)

	TIMSS 2003		TIMSS 2011		TIMSS 2015		Changes			
	$\%_{02}$	(SE)	$\%_{11}$	(SE)	$\%_{15}$	(SE)	$\%_{15}–\%_{02}$	(SE)	$\%_{15}–\%_{11}$	(SE)
Content	61,4	(4,6)	72,5	(3,4)	84,2	(3,0)	22,9*	(3,0)	11,7	(4,0)
Pedagogy/instruction	42,9	(4,0)	49,5	(3,8)	58,2	(3,6)	15,3*	(3,6)	8,7	(4,5)
Curriculum	59,2	(3,7)	70,5	(3,7)	86,4	(2,4)	27,2*	(2,4)	15,9*	(3,6)
Critical thinking or problem solving	57,9	(4,2)	51,2	(3,7)	55,9	(3,2)	–2,0	(3,2)	4,7	(4,2)
Assessment	76,6	(3,4)	68,8	(3,7)	73,1	(2,7)	–3,5	(2,7)	4,3	(3,9)
Integrating information technologies	34,2	(4,0)	35,2	(3,4)	45,5	(3,5)	11,2*	(3,5)	10,3	(4,4)

* = significant change ($p < 0.05$)

Table 11.8 Participation in professional development in mathematics over two years in 2002, 2011, and 2015 by topic and qualification profile (percentages of Grade 8/9 learners in percent)

	Degree on university level, studied mathematics or mathematics education as			Degree below university level, studied mathematics or mathematics education			Degree on university level, studied other subject as a major			Degree below university level, studied other subject as a major		
		%	SE		%	SE		%	SE		%	SE
2015	↑ Curriculum	86,1	(3,1)	↑ Curriculum	92,7	(3,5)	↑ Curriculum	77,8	(10,4)	↑ Curriculum	72,7	(16,7)
	↑ Content	85,5	(3,8)	↑ Content	89,2	(4,7)	↑ Content	74,7	(10,9)	Content	53,7	(17,5)
	Assessment	73,9	(3,7)	Assessment	77,0	(6,5)	↓ Assessment	69,3	(10,1)	↓ Assessment	43,3	(14,4)
	Pedagogy	59,2	(5,0)	↑ Pedagogy	66,9	(6,4)	Critical thinking	61,5	(12,4)	Critical thinking	40,1	(14,1)
	Critical thinking	54,8	(4,1)	Critical thinking	59,1	(7,0)	IT	56,8	(12,3)	IT	22,6	(12,1)
	IT	43,1	(4,4)	IT	50,8	(7,1)	Pedagogy	46,1	(12,5)	Pedagogy	17,9	(10,0)
2011	Assessment	74,3	(8,3)	Assessment	78,1	(4,4)	↑ Assessment	80,2	(7,3)	↑ Assessment	71,6	(13,7)
	Curriculum	66,9	(8,5)	Content	63,3	(5,6)	Curriculum	59,8	(8,9)	Curriculum	50,0	(14,2)
	↓ Content	60,3	(9,2)	Critical thinking	62,6	(4,5)	↓ Content	54,8	(11,5)	↓ Content	46,8	(14,2)
	↑ Critical thinking	54,8	(8,5)	↓ Curriculum	57,3	(5,3)	Critical thinking	54,7	(10,8)	Critical thinking	40,2	(14,3)
	Pedagogy	50,5	(9,4)	Pedagogy	41,8	(5,2)	↑ IT	42,3	(10,7)	↑ Pedagogy	35,5	(13,0)
	IT	33,7	(7,9)	IT	32,5	(4,8)	Pedagogy	33,1	(9,8)	IT	22,5	(10,9)
2003	Content	76,9	(4,9)	Assessment	75,8	(5,5)	Content	70,9	(13,8)	Content	82,7	(10,0)
	Curriculum	72,9	(5,5)	Curriculum	74,7	(6,2)	Assessment	62,7	(13,7)	Curriculum	69,2	(13,3)
	Assessment	66,6	(5,2)	Content	70,0	(6,1)	Curriculum	54,7	(13,8)	Assessment	65,7	(14,3)
	Pedagogy	51,5	(5,4)	Critical thinking	66,4	(5,6)	Critical thinking	41,4	(13,0)	Critical thinking	31,9	(14,4)
	Critical thinking	43,5	(5,1)	Pedagogy	54,2	(6,8)	Pedagogy	34,6	(12,0)	IT	27,0	(13,3)
	IT	37,3	(4,8)	IT	39,4	(6,5)	IT	23,7	(10,8)	Pedagogy	21,8	(12,8)

↑↓ Relative change in positioning in addition to a notable gain/lost in percentage points (min. 10%)

university degree showed significant lower mathematics achievement than students taught by fully qualified teachers. In addition it can be found that, when looking at the results for model 1, the used teacher qualification indicators become less valuable to explain differences in student achievement over time: Whereas in 2003 about 12% of the total variance could be explained by the teacher qualification indicator, it was only about 3 and 2% respectively, in 2011 and 2015.

In Model 4 we tested for differential effectiveness, as previous research suggested that teacher qualification may not have an overall effect but rather play a role dependent on the school context. Overall there is more evidence to not reject the null hypothesis and conclude that, on average in South Africa differences in formal qualification profiles do also not have context specific statistically significant, predictive differential relationships with mathematics achievement of Grade 8/9 students, net of all else, in all three study cycles. An exception is TIMSS 2015 where students taught by out-of-field teachers with a university degree showed significantly higher mathematics achievement than students taught by fully qualified mathematics teachers with a university degree. The significant negative value of the interaction coefficient suggests in addition that for these students' structural disadvantages resulting from high proportions of students from families living in poverty are less stark. With caution this may be interpreted as an indicator for differential effectiveness suggesting that teachers with these qualification profiles, in comparison to their colleagues with other qualification profiles, are more effective in reducing inequity in mathematics achievement resulting from structural differences between schools.

Over all study cycles student age and self-concept as well as the number of students living under poverty conditions, are significantly associated with achievement in mathematics, whereas all other covariates show only significant relationships in some or none of the study cycles. Another important finding is the changes in the distribution of variance across levels and the respective variance explained over time.

Table 11.9 Models of multilevel regressions analysis with math proficiency as DV by controlling for teacher, student, and class characteristics

	Model 1			Model 2			Model 3			Model 4		
	TIMSS 2003	**TIMSS 2011**	**TIMSS 2015**	**TIMSS 2003**	**TIMSS 2011**	**TIMSS 2015**	**TIMSS 2003**	**TIMSS 2011**	**TIMSS 2015**	**TIMSS 2003**	**TIMSS 2011**	**TIMSS 2015**
Fixed Effects												
Constant	306,2 (20,0)	372,1 (8,3)	385,7 (6,4)	250,9 (42,5)	318,9 (51,4)	408,8 (41,2)	260,3 (43,2)	303,7 (55,3)	416,9 (43,8)	222,0 (53,3)	267,4 (57,1)	378,4 (44,2)
Class Level												
Teacher Characteristics												
Non-University Ed & Major (Mathematics) [1]	**-63,8** (21,6)	**-31,7** (14,1)	**-20,9** (10,3)	-16,0 (10,2)	-7,7 (11,7)	-7,5 (9,7)	-10,0 (10,8)	-7,6 (8,1)	-2,8 (11,9)	46,3 (42,7)	45,8 (30,6)	86,2 (70,1)
University Education: Other Spec.[1]	-0,6 (26,1)	-21,9 (23,5)	-19,6 (15,4)	-6,5 (10,7)	8,9 (16,4)	0,4 (10,9)	-2,6 (12,0)	10,5 (14,3)	5,6 (11,7)	18,4 (43,9)	111,7 (71,2)	**88,3** (41,6)
Non-University Ed & Other Spec.[1]	**-62,1** (22,6)	8,2 (38,1)	-20,7 (20,6)	**-28,9** (13,8)	-25,6 (19,7)	-6,5 (11,1)	-23,8 (15,3)	-23,5 (21,2)	-13,1 (12,9)	52,9 (84,4)	60,7 (74,5)	-60,0 (52,7)
Experience; 6-10 y.[2]							-9,4 (9,7)	13,9 (13,0)	5,7 (9,2)	-4,7 (10,2)	12,7 (12,2)	3,2 (8,4)
Experience: 11-20 y.[2]							-2,8 (9,1)	7,0 (10,4)	-7,3 (9,6)	2,0 (9,4)	4,5 (8,9)	-8,2 (9,6)
Experience: at least 20 y.[2]							-10,4 (13,6)	-16,1 (10,3)	-10,1 (10,6)	-10,9 (14,0)	-15,1 (9,3)	-9,3 (9,4)
Intensive participation in PD[3]							-1,8 (1,8)	0,1 (1,8)	-2,2 (1,7)	-1,7 (1,8)	0,2 (1,5)	-2,2 (1,5)
Sex: female[4]							1,6 (6,9)	8,1 (8,7)	7,0 (7,6)	2,3 (7,1)	5,3 (7,3)	7,3 (7,4)
Class composition												
Class size [(a)]				**-1,2** (0,4)	-0,2 (0,2)	-0,4 (0,3)	**-1,2** (0,4)	-0,1 (0,2)	**-0,5** (0,3)	**-1,1** (0,4)	-0,1 (0,2)	-0,5 (0,3)
Small proportion of students from families living in poverty[(as)14]				**254,0** (33,0)	**292,4** (37,6)	**303,4** (49,9)	**237,3** (31,9)	**315,2** (37,0)	**312,1** (48,3)	**291,4** (57,3)	**373,7** (45,5)	**364,6** (45,3)
High proportion of students speaking the test language at home [(as)]				**113,5** (17,1)	**42,6** (20,9)	18,7 (21,2)	**123,5** (18,3)	**38,6** (19,2)	14,7 (16,0)	**116,6** (18,7)	**34,9** (18,7)	17,8 (15,7)
Interactions												
Non-University Ed & Major (Mathematics) * Small proportion of students from families living in poverty										-105,7 (69,7)	-81,4 (50,5)	-138,1 (110,8)
University Education: Other Spec. * Small proportion of students from families living in poverty										-34,0 (65,8)	-153,3 (100,2)	**-131,3** (59,6)
Non-University Ed & Other Spec.* Small proportion of students from families living in poverty										-137,8 (154,6)	-119,3 (105,5)	67,7 (71,4)

(continued)

Table 11.9 (continued)

Student level																					
Sex: female[6]				-4,7	(5,6)	-6,0	(4,3)	**-9,0**	(3,3)	-4,9	(5,6)	-6,5	(4,4)	**-8,8**	(3,4)	-4,8	(5,6)	-6,6	(4,5)	**-8,7**	(3,4)
Age				**-10,5**	(2,1)	**-11,6**	(2,4)	**-15,5**	(1,8)	**-10,4**	(2,1)	**-12,0**	(2,4)	**-15,5**	(1,8)	**-10,4**	(2,1)	**-12,0**	(2,4)	**-15,5**	(1,8)
High self-concept[7]				**22,5**	(3,3)	**19,3**	(3,0)	**16,1**	(2,8)	**22,8**	(3,5)	**19,7**	(3,2)	**16,6**	(3,0)	**22,9**	(3,5)	**19,6**	(3,2)	**16,6**	(2,9)
Speaks (almost) always test language at home[8]				11,5	(9,6)	7,3	(5,8)	**12,3**	(4,3)	10,9	(9,8)	8,1	(5,9)	**12,4**	(4,4)	10,9	(9,8)	8,1	(5,9)	**12,3**	(4,4)
Parents with max. upper second. Education[9]				5,3	(6,8)	-4,4	(4,8)	-2,9	(4,2)	5,6	(7,2)	-4,6	(5,2)	-2,4	(4,4)	5,6	(7,2)	-4,4	(5,2)	-2,3	(4,4)
Parents with min. University degree[9]				**18,5**	(9,1)	1,8	(6,1)	-2,0	(5,3)	**20,5**	(9,6)	1,5	(6,4)	-1,4	(5,7)	**20,3**	(9,6)	1,7	(6,4)	-1,3	(5,7)
Own study desk at home[10]				1,2	(5,9)	-3,4	(3,9)	-4,9	(3,5)	3,5	(6,2)	-3,3	(4,1)	-5,5	(3,6)	3,3	(6,2)	-3,4	(4,1)	-5,6	(3,6)
More than 25 books at home[11]				-0,3	(6,9)	-0,3	(4,7)	3,6	(4,0)	-0,8	(7,6)	0,1	(4,8)	3,4	(4,0)	-1,1	(7,6)	0,1	(4,8)	3,4	(4,0)
Electricity at home[12]				8,4	(8,0)	6,4	(8,8)	5,1	(7,2)	7,2	(8,0)	8,5	(9,1)	4,9	(7,8)	7,7	(8,1)	9,3	(9,0)	5,3	(7,9)
Running tap water at home[13]				7,4	(6,0)	**10,2**	(5,6)	4,8	(4,7)	6,7	(7,1)	9,6	(5,7)	4,5	(4,9)	6,9	(7,1)	9,5	(5,7)	4,5	(4,9)
Variance																					
Between schools (T03: 67,6 %; T11: 61,4%; T15: 60,0%)	17,6	5,4	3,4	82,6		76,4		60,6		84,2		80,6		60,7		84,8		82		64	
Within schools (T03: 32,4 %; T11: 38,6%; T15: 40,0%)	0	0	0	8,8		13		16,2		8,8		13		16,2		8,8		13		16,2	
Total	11,9	3,3	2	58,7		51,9		42,8		59,8		54,5		42,9		60,2		55,4		44,9	

Note DV = Global scale of proficiency in mathematics. Significant coefficients in bold ($p < .05$). Standard errors in brackets
1 Reference group: University education & main study area/major was mathematics (t), 2 Reference group: Experience: 0 to 5 years (t). 3 Values between 0 to 5 were higher value shows intensity in covering a variety of topics in mathematics in formal professional development activities over the last two years (t); 4 Reference group: male (t). 5 Reference group: 20 % of classes with the highest number of students from disadvantaged backgrounds (as) see 14. 6 Reference group: male (s); 7 Values between 0 and 3, were 3 indicates a high positive response on a scale comprising 3 items on (e.g. *I learn things quickly in mathematics*); 8 student speaks never or sometimes test language at home (Afrikaans & English); 9 Highest Education level of both parents Lower Secondary Education (s); 10 no study desk at home (s); 11 less than 25 books at home (s); 12 no electricity at home (s); 13 no running tap water at home (s); 14 Aggregated student information on average availablity of electricity, water, a stuy room, a study desk and min. 25 books at home
(a) aggregated nr. of students; (s) student reports; (t) teacher reports; (as) aggregated student responses

Whereas in 2003 about 68% of the variance was associated with between school differences, this reduced to 60% in 2015. In association it can be found that the school level predictors reduce their explanatory power over time, whereas student level covariates become more valuable in predicting mathematics achievement.

11.11 Summary of Findings

We designed our study to investigate to what extent teacher qualification profiles and respective allocation have changed over the last 15 years in South Africa and to what extent these profiles were associated with different participation in professional development activities and students mathematics achievement. The results support a few substantively important conclusions.

South Africa has made substantial progress in uplifting teacher's formal qualification levels.

We show that between 2003 and 2015 the number of students taught by fully qualified teachers in mathematics increased significantly from about 24% to about 62%. This finding corresponds with the findings presented by Arends (2013) and Reddy et. al. (2016).

South Africa has made substantial progress in reducing structural inequality within its education system.

We show that over time the variance associated with between school levels substantially decreased while at the same time student-level covariates become more valuable in predicting mathematics achievement. We interpret this finding as a substantial reduction of structural inequality within South Africa'seducation system.

Out-of-field teaching is still a common phenomenon and unevenly distributed.

We show that even though the number of students taught by out-of-field teacher also significantly decreased, in 2015 about 14% of all Grade 9 students were taught by teachers in mathematics that did not specialize in mathematics throughout their initial teacher training. These fully qualified teachers are more likely to be found in urban schools with comparatively privileged students. This finding partially corresponds with qualitative research presented by Du Plessis et al. (2014), Onwu and Sehole (2010), and Mampane (2009). For 2015 we found that learners schooled in urban locations were significantly more often taught by fully qualified teachers than their peers going to school in medium size cities, towns, or remote rural location. However against our expectation, we do not find a systematic disadvantage of schools in remote rural areas, possibly a result of a well targeted allocation policy in correspondence with a general higher number of fully qualified teachers.

Teachers with different qualification profiles do not differ in their usage of professional development opportunities.

In accordance with policy we do find an increase in teacher qualification in professional development. We do not find any support for the theoretical assumptions of different patterns of participation in professional development depending on qualification profiles. We rather find general trends in favor of certain topics. Qualitative research on usages and systematic research on offerings would be needed to further understand this pattern. Initial observations are that with the introduction of the new Curriculum and Assessment Policy Statement (CAPS) in 2011, assessment took on increased importance. A further observation is that while problem solving was a concerted focus in the 1980s, the offering and participation in this areas is relatively less. The demise of reform mathematics, with a focus on problem solving, is to be further explored, especially in the light of our qualified mathematics teachers not making a significant difference to learner achievement.

Teacher's formal level of education is in general not significantly associated with student's mathematics achievement. If at all, differential advantages are rather found for formal qualification than specialization.

Contrary to expectation, we find no significant relationship between teacher qualification and student's mathematics achievement. This might be because in contrast to previous studies we chose a multivariate multilevel approach to study the relationship. Whereas in our bivariate control model we also find a significant relationship, these results disappear when we control for composition and student covariates to control for different allocation patterns. Based on this finding we argue that previously reported differences can probably be attributed to specification errors, not taking systematic differences in the allocation of teachers to different contexts into account. In accordance with Shepherd (2013) and Spaull (2011) we do find a differential relationship for the TIMSS 2015 but in favor of out-of-field teachers with a university degree. This finding is surprising given the theoretical framework of the study according to which we assumed that fully qualified mathematics teachers with a university degree should have been most effective in producing both higher achieving students and reducing inequality. For interpretation it should be noted that this finding is only of relevance for a comparatively small number of students who are taught by teachers with such a qualification profile (10%). It might well be that this finding is biased as a result of an omitted variable.

References

Adler, J., Pornara, C., Taylor, D., Thorne, B., & Moletsane, G. (2013). Mathematics and science teacher education in South Africa: A review of research, policy and practice in times of change. *African Journal of Research in Mathematics, Science and Technology Education, 13*, 28–46.

Arend, F. (2011). Teacher shortages. The need for more reliable information on school level. In HSRC (Eds.), *RESDI*. http://www.hsrc.ac.za/uploads/pageContent/2702/RESDI%20newsletter,%20November%202010%20issue.pdf

Arends, F. (2013). *The good teacher: What teachers need to teach well*. http://www.hsrc.ac.za/en/review/hsrc-review-may-2013/the-good-teacher-what-teachers-need-to-teach-well

Arends, F., Winnaar, L., & Mosimege, M. (2017). Teacher classroom practices and Mathematics performance in South African schools: A reflection on TIMSS 2011. *South African Journal of Education, 37*(3), 1–11. https://doi.org/10.15700/saje.v37n3a1362

Beck, E., et al. (2008). *Adaptive Lehrkompetenz. Analyse und Struktur, Veränderbarkeit und Wirkung handlungssteuernden Lehrerwissens*. Waxmann.

Carnoy, M., & Arends, F. (2012). Explaining mathematics achievement gains in Botswana and South Africa. *Prospects, 42*(4), 453–468. https://doi.org/10.1007/s11125-012-9246-6

Carnoy, M., Chisholm, L., et al. (2008). *Towards understanding student academic performance in South Africa: A pilot study of grade 6 mathematics lessons in South Africa.* HSRC. www.hsrc.ac.za/en/research-data/ktree-doc/1392

Chisholm, L., Hoadley, U., wa Kivulu, M., Brookes, H., Prinsloo, C., Kgobe, A., Mosia, D., Narsee, H., & Rule, S. (2005). *Educator workload in South Africa*. HSRC.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources, 41*(4), 778–820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2012). Teacher credentials and student achievement in high school. *Journal of Human Resources, 45*(3), 655–681.

Coldwell, M. (2017). Exploring the influence of professional development on teacher careers: A path model approach. *Teaching and Teacher Education, 61*, 189–198.

Council on Higher Education. (2007). *The national policy framework for teacher education and development in South Africa*. Retrieved on 23 April 2018 from: http://www.che.ac.za/media_and_publications/frameworks-criteria/national-policy-framework-teacher-education-and

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1). http://epaa.asu.edu/ojs/article/view/392

DBE (2014) = Department of Basic Education (DBE). Department of Basic Education Education for All (EFA) 2014 Country Progress Report. Pretoria: DBE. https://www.education.gov.za/Portals/0/Documents/Reports/2014%20Education%20For%20All%20(EFA)%20Country%20Progress%20Report.pdf?ver=2015-02-18-130341-697

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199.

Desimone, L. M., Smith, T., & Ueno, K. (2006). Are teachers who need sustained: Content-focused professional development getting it? An Administrator's Dilemma. *Educational Administration Quarterly, 42*(2), 179–215.

Dee, T. S., & Cohodes, S. R. (2008). Out-of-field teachers and student achievement: Evidence from "matched-pairs" comparisons. *Public Finance Review, 36*(7), 7–32.

DHET. (2015) Department of Higher Education and Training. (2015). The minimum requirements for teacher education qualifications. *Government Gazette No. 38487, Vol. 596*. Retrieved from http://www.gov.za/sites/www.gov.za/files/38487_gon111.pdf

Du Plessis, A. E. (2013). *Understanding the out-of-field teaching experience.* A thesis submitted for the degree of Doctor of Philosophy at the University of Queensland. Online: http://espace.library.uq.edu.au/view/UQ:330372/s4245616_phd_submission.pdf

Du Plessis, A. E., Gillies, R. M., & Carroll, A. (2014). Out-of-field teaching and professional development: A transnational investigation across Australia and South Africa. *International Journal of Educational Research, 66*, 90–102. https://doi.org/10.1016/j.ijer.2014.03.002

Education Labour Relations Council (ELRC). (2003). *Integrated quality management system (IQMS) for school-based educators*. ELRC.

Goldhaber, D. D. (2002). The mystery of good teaching. *Education next, 2*(1), 50–55.

Goldhaber, D. D., & Brewer, D. J. (1996). Evaluating the effect of teacher degree level on educational performance. *Developments in School Finance*, 199–210. http://nces.ed.gov/pubs97/97535l.pdf

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*, 129–146.

Gustafsson, M. (2007, March). Using the hierarchical linear model to understand school production in South Africa. *South African Journal of Economics, 75*(1), 84–98.

Glewwe, et al. (2011). http://www.nber.org/papers/w17554.pdf

Hofmeyr, J., & Hall, G. (1995). *The National Teacher Education Audit: Synthesis report*. Johannesburg: Edupol, National Business Initiative

Hobbs, L. (2013). Teaching 'out-of-field' as a boundary-crossing event: Factors shaping teacher identity. *International Journal of Science and Mathematics Education, 11*, 271–297.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications.* Routledge.

Huberman, M. (1989). The professional life cycles of teachers. *Teacher College Record, 91*(1), 31–58.

Huberman, M. (1993). *The lives of teachers.* Cassell.

Jansen, J. & Taylor, N. (2003, October). *Educational change in South Africa 1994–2003: Case studies in large-scale education reform.* Country Studies – Education Reform and Management Publication Series, 11(1).

Kasper, D., Schulz-Heidorf, K., & Schwippert, K. (2018). %SURVEYHLM: A SAS® macro for multilevel analysis with large-scale educational assessment data. *Manuscript under review.*

Kunene, L. L. Z. (2013). *Classroomlevel factors affecting mathematics achievement: A comparative study between South Africa and Australia using TIMSS 2003.* Dissertation (MEd), University of Pretoria. https://repository.up.ac.za/handle/2263/25819

McConney, A., & Price, A. (2009). Teaching out-of-field in Western Australia: Australian. *Journal of Teacher Education, 34*(6), 86–100.

May, P. (2006). Englisch-Hörverstehen am Ende der Grundschulzeit. In W. Bos & M. Pietsch (Eds.), *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen* (pp. 203–224). Waxmann.

Mampane, S. T. (2009). *How school governing bodies understand and implement changes in legislation with respect to the selection and appointment of teachers* (PhD Thesis). University of Pretoria. https://repository.up.ac.za/handle/2263/28176

Monk, D. H., & King, J. (1994). Multi-level teacher resource effects on pupil performance in secondary Mathematics and Science: The role of teacher subject-matter preparation. In R. Ehrenberg (Ed.), *Contemporary policy issues: Choices and consequences in education* (pp. 29–58). ILR Press.

Morris, M., Chrispeels, J., & Burke, P. (2003). The power of two: Linking external with internal teachers' professional development. *The Phi Delta Kappa International, 84*(10), 764–767.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

Mullis, I. V. S., Martin, M. O., Goh, S., & Cotter, K. (Eds.). (2016). *TIMSS 2015 encyclopedia: Education policy and curriculum in mathematics and science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2015/encyclopedia/

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research.* Oxford University Press.

Onwu, G. O. M., & Sehoole, C. (2010). *Why Teachers matter: Policy issues in the professional development of teachers in South Africa* (pp. 121–136). http://aadcice.hiroshima-u.ac.jp/e/publications/sosho4_2-10.pdf

Prinsloo, C., & Rogers, S. (2013). The missing link: language skills crucial to mathematics and science. *HSRC Review, 11*(2), 26–27. Available at http://www.hsrc.ac.za/en/review/hsrc-review-may-2013/the-missing-link-language-skills-crucial-to-mathematics-and-science. Accessed 7 July 2017.

Poti, J., Mutsvangwa, A., & Hove, M. (2014). Teacher retention and quality education: Impact of rural incentives in North-West, South Africa. *Journal of Social Sciences, 5*, 792–806. https://doi.org/10.5901/mjss.2014.v5n27p792

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7 for Windows* [Computer software]. Scientific Software International, Inc.

Richter, D., Kuhl, P., Reimers, H., & Pant, H. A. (2012). Aspekte der Aus- und Fortbildung von Lehrkräften in der Primarstufe. In P. Stanat, H. A. Pant, K. Böhme, & D. Richter (Eds.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (pp. 237–250). Waxmann.

Richter, D., Kuhl, P., Haag, N., & Pant, H. A. (2013). Aspekte der Aus- und Fortbildung von Mathematik- und Naturwissenschaftslehrkräften im Ländervergleich. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (pp. 367–390). Waxmann.

Reeves, C., & Robinson, M. (2010). Am I 'qualified' to teach? The implications of a changing school system for criteria for teacher qualifications. *Journal of Education, 50,* 1–33. http://joe.ukzn.ac.za/Libraries/No_50_2010/Am_I_qualified_to_teach_The_implications_of_a_changing_school_system_for_criteria_for_teacher_qualifications.sflb.ashx

Reddy, V., Visser, M., Winnaar, L., Arends, F., Juan, A., Prinsloo, C. H. (2016). *TIMSS 2015: Highlights of mathematics and science achievement of grade 9 South African Learners.* Human Sciences Research Council. www.hsrc.ac.za/en/research-data/ktree-doc/17642

Reddy, V., Prinsloo, C., Arends, F., Visser, M., Winnaar, L., Feza, N., Rogers, S., Janse van Rensburg, D., Juan, A., Mthethwa, M., Ngema, M., & Maja, M. (2015). *Highlights from TIMSS 2011: The South African perspective.*

Reddy, V., Zuze, T. L., Visser, M., Winnaar, L., Juan, A., Prinsloo, C. H., Arends, F., & Rogers, S. (2015). *Beyond benchmarks: What twenty years of TIMSS data tell us about South African Education.* HSRC Press. http://pub.iea.nl/fileadmin/user_upload/Publications/National_reports/TIMSS_2011_report_SouthAfrica.pdf

Reddy, V. (Ed.). (2006). *Mathematics and science achievement at South African schools in TIMSS 2003.* HSRC Press.

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*, 116–126.

Reddy, V., Van der Berg, S., Janse van Rensburg, D., & Taylor, S. (2012). Educational outcomes: Pathways and performance in South African high schools. *South African Journal of Science, 108*(3/4): Art. 620, 8 pages. https://doi.org/10.4102/sajs.v108i3/4.620

Robinson, R., & Carrington, S. (2002). Professional development for inclusive schooling. *International Journal of Educational Management, 16*(5), 239–247.

Shepherd, D. L. (2013, June 23–25). *The impact of teacher subject knowledge on learner performance in South Africa: A within-pupil across-subject approach.* Paper presented at the International Workshop on Applied Economics of Education, Italy. Available at http://www.iwaee.org/papers%20sito%202013/Shepherd.pdf. Accessed 29 June 2018.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15,* 4–14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1–23.

Sayed, Y. (2004). The case of teacher education in Post-Apartheid South Africa: Politics and priorities. In L. Chisholm (Ed.), *Changing class: Education and social change in Post-Apartheid South Africa.* HSRC Press.

Spaull, N. (2011). *A preliminary analysis of SACMEQ III South Africa* (No. 09/2013). Working Papers. Stellenbosch University, Department of Economics

Spaull, N. (2012). *Poverty & privilege: Primary school inequality in South Africa.* (Working Papers 13/2012). Stellenbosch University, Department of Economics.

Steyn, G. M. (2011). Continuing professional development in South African schools: Staff perceptions and the role of principals. *Journal of Social Sciences, 28*(1), 43–53. http://uir.unisa.ac.za/handle/10500/13410

Taylor, N., & Vinjevoild, P. (1999). *Getting learning right: Report of the President's Education Initiative Research Project*. Retrieved from http://jet.org.za/publications/books/getting-learning-right

Westheimer, J. (2008). Learning among colleagues: Teacher community and the shared enterprise of education. In M. Cochran-Smith, S. Feiman-Nemser, J. McIntyre, & K. E. Demers (Eds.), *Handbook of research on teacher education* (pp. 756–782). Routledge.

Van der Berg, S. (2008). How effective are poor schools? Poverty and educational outcomes in South Africa. *Studies in Educational Evaluation, 34*(3), 145–154. https://doi.org/10.1016/j.stueduc.2008.07.005

Visser, M., Juan, A., & Feza N. (2015). Home and school resources as predictors of mathematics performance in South Africa. *South African Journal of Education, 35*(1): Art. # 1010, 10 pages. https://doi.org/10.15700/201503062354

Winnaar, L. D., Frempong, G., & Blignaut, R. (2015). Understanding school effects in South Africa using multilevel analysis: findings from TIMSS 2011. *Electronic Journal of Research in Educational Psychology, 13*(1): 151–170. https://doi.org/10.14204/ejrep.35.13116

Zuzovsky, R. (2009). Teachers' qualifications and their impact on student achievement: Findings from TIMSS 2003 data for Israel. In M. V. Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (IERI Monograph Series, Vol. 2, pp. 37–62). IEA-ETS Research Institute. www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02.pdf

Heike Wendt, PhD., is Professor for Education Research, Institute of Education Research and Teacher Education, Faculty of Environmental, Regional, and Educational Sciences, Graz University, Austria.

Dr. Daniel Kasper is with the Department of Evaluation of Educational Systems at the University of Hamburg, Germany. His areas of research include large-scale assessment, psychometrics, and statistics.

Caroline Long, Ph.D., is a Professor at the Department of Childhood Education, University of Johannesburg, South Africa. Her research areas span curriculum, mathematics education, and assessment and evaluation, in particular the relationship of measurement to assessment within a Rasch measurement framework.

Chapter 12
Revisiting the Relationship Between Science Teaching Practice and Scientific Literacy: Multi-level Analysis Using PISA

Hyesun You

Abstract Growing evidence from recent curriculum documents and previous research suggests that inquiry-based science teaching practices promote students' conceptual understanding, level of achievement, and motivation to learn. However, some researchers have questioned whether inquiry-based learning is the best learning method and have claimed that direct instruction is more efficient and equally effective for student performance. To contribute to this debate, the current study, drawing on data from the Program for International Student Assessment 2015, used a multivariate multilevel method to examine the relationship between the scientific literacy of 5712 American students from 177 schools and two teaching practices: inquiry-based teaching and direct instruction. The results of multilevel modeling, after controlling for student- and school-level variables, revealed that inquiry-based teaching was significantly negatively related to scientific literacy, whereas direct instruction was significantly positively related to scientific literacy. The findings of this study can help achieve a comprehensive understanding of science teaching practices and students' performance on an international test, and this understanding can provide insights for future teaching strategies.

Keywords Inquiry-based teaching · Direct instruction · Scientific literacy · PISA · Hierarchical linear model

12.1 Introduction

Since the first international study of science learning outcomes was conducted between 1966 and 1973, American students have fared poorly in science assessments, lagging behind students from other countries (Medrich & Griffith, 1992). The science performance of American 15-year-olds in the results of the Program for International Student Assessment (PISA), an international test, has not improved much since 2000.

H. You (✉)
University of Texas, Austin, TX, USA
e-mail: hyesun-you@uiowa.edu

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_12

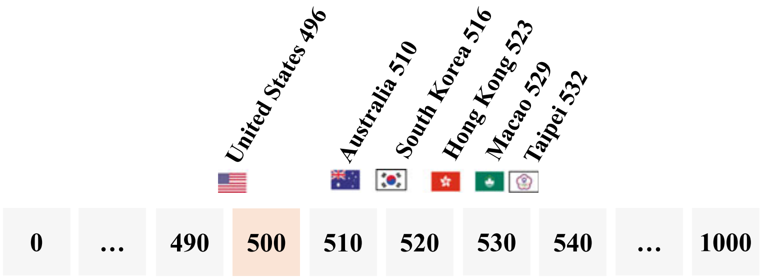


Fig. 12.1 How far behind are the U.S. students in science? (from the PISA 2015 data)

The latest 2018 results of PISA showed some improvement in U.S. science performance, but they were still near the Organisation for Economic Co-operation and Development (OECD) average. This indicates that American students continue to lag behind their peers in some regions in East Asia and Europe (Fig. 12.1). Moreover, the achievement gap in science between high and low performers is widening. The top quarter of American students have improved their performance on the exam since 2000, but the bottom 10th percentile's performance has become worse (OECD, 2019).

Academic improvement in science has been the focus of bipartisan education legislation for decades, costing many billions of dollars and resulting in a string of national programs (e.g., No Child Left Behind of 2002, Race to the Top of 2009, NGSS of 2013, and the Every Student Succeeds Act). Many reformers have wanted to raise standards and help students compete with their peers across the globe and to improve schools and teachers through policy and educational framework initiatives, but the PISA test scores cast doubt on these efforts.

Since the publication of the National Science Education Standards (NSES) by the National Research Council (NRC) in 1996, scientific inquiry has been emphasized in K-12 science education in the United States, and it has influenced multilayered teaching practices and students' science learning. Other U.S. national reform documents, including the *Framework for K-12 Science Education* (Framework; NRC, 2012) and the *Next Generation of Science Standards* (NGSS; NGSS Lead States, 2013), stated the importance of inquiry-based teaching (IBT) as follows: teachers should allow students to reveal their knowledge by "doing" a task using that knowledge. As such, IBT practices (learning to do science) stand in sharp contrast to traditional instruction (teaching about science), where teachers lecture and direct students through step-by-step activities. Along with national-reform-oriented standards, many educational programs, centers, and reform initiatives have been established to help promote hands-on inquiry in science teaching and learning in schools throughout the country.

An extensive body of research has shown that reform-oriented, effective IBT practices can be mediated to improve student learning (e.g., Adak, 2017; Schuster et al., 2018). Nevertheless, as IBT has become dominant, researchers have often expressed concerns about the pedagogical value of the inquiry-based approach and students' unfavorable academic outcomes (e.g., Edelson et al., 1999; Kirschner et al., 2006). For example, a recently published article by Liou (2020), on the case of Taiwanese students, supported the argument that inquiry-based instructional practice may have a significant negative impact on PISA's scientific literacy.

The current study revisits this long-standing debate about IBT and direct instruction (DI) to determine which teaching practice is better for scientific literacy by using the 2015 PISA dataset, and by taking account of many other confounding factors closely related to science achievement.

12.2 Literature Review

12.2.1 Benefits and Challenges of DI and IBT

Two dominant teaching models, DI and IBT, represent distinct perspectives on instructional approaches, which allows us to present some advantages and disadvantages that must be weighed against one another. DI is a form of teacher-led instruction that presents new information in a sequentially organized fashion with a structured curriculum. DI is the best way of providing detailed information or encouraging systematic skill acquisition and is appropriate when memorization and immediate recall is desired (Dean & Kuhn, 2007). Esler and Sciortino (1991) argued that the practice is especially effective for students at the early elementary level and for low-achieving students at the secondary level. However, this practice is often criticized for preparing students for the next grade level and on in-school success rather than on helping students learn more authentic ways for meaningful learning. Also, students can be passive learners in DI lessons, and the information delivered to them can be lost over time if it is not used or applied (Cobern et al., 2010).

Since the 1990s, science standards and preeminent science educators have advocated inquiry as a more preferred method of teaching science. IBT is encouraged because science is fundamentally a question-driven exploration and inference process, and students need active participation opportunities to derive their own questions, develop their own methods, carry out their own investigations, draw their own conclusions through scientific inquiry, and construct new scientific knowledge (NRC, 1996). Despite this apparent consensus about IBT, some have made the claim that IBT is not the best teaching method for academic success. Critics of IBT have argued that a minimally guided approach in IBT does not provide sufficient structure to help students learn the important concepts and procedures of science (Kirschner et al., 2006).

Edelson et al. (1999) indicated several challenges to the process of implementing inquiry-based learning and revealed five main drawbacks to its use as a teaching method: (1) motivation—when students are not sufficiently motivated, they either fail to participate in inquiry activities, or they participate in a disengaged manner that does not support learning; (2) accessibility of investigation techniques—if students are not able to master scientific investigation techniques such as data collection and analysis, then they cannot conduct investigations and interpret the results; (3) background knowledge—if students lack content knowledge related to inquiry practices, then they will be unable to complete meaningful investigations; (4) management of extended activities—if they are unable to organize their work and manage an extended process, students cannot engage in open-ended inquiry; and (5) the practical constraints of the learning context—a failure to work within the available technology or fit within the existing schedule of a school will lead to the failure of IBT. In particular, other scholars have indicated that IBT is difficult for low-performing students with limited prior knowledge. Their lower level of background knowledge discourages them from playing a proactive role in the learning process for new knowledge construction.

12.2.2 Review of Impacts of Teaching Practices (Inquiry Versus Direct) on Science Achievement

Although the debate between IBT and traditional instruction has continued, a variety of studies have been conducted on the effectiveness of using IBT and direct teaching practices in an experimental or quasi-experimental design in which student performance is compared across two groups of students: a control group taught in a lecture-based manner and a treatment group taught with some form of IBT. This research has proceeded under the hypothesis that students exposed to IBT tend to perform better than students in the control conditions. In fact, many empirical studies have proven the effectiveness of IBT on science performance (e.g., Böttcher & Meisert, 2013; Minner et al., 2010; Schroeder et al., 2007; Stender et al., 2018; Wolf & Fraser, 2008). Furtak et al.'s meta-analysis study (2012) also supported IBT with a medium overall effect size of 0.5 across a total of 37 studies.

Although the research showed the effectiveness of the IBT approach, some researchers have expressed the critics' view that the inquiry-learning environment does not seem ideal for prompting students' conceptual development. Even some literature argued that the best methodology for teaching students about the nature of science is explicit instruction for novice-to-intermediate learners (Kirschner et al., 2006). Klahr and Nigam (2004) made a case for the superiority of DI over discovery learning. They compared the abilities of 112 third- and fourth-grade students to design scientific investigations (e.g., designing and interpreting experiments and applying basic skills to more authentic reasoning) under two conditions, DI and discovery-based learning. The results showed that many students learned much better from DI than from discovery learning about scientific judgments and experimental design.

Some Trends in International Mathematics and Science Study (TIMSS) assessments have revealed a negative or curvilinear relationship between IBT and student achievement in specific countries. Kaya and Rice (2010) found that science scores were negatively related to IBT in Australia and the United States and only positively related in Singapore in the TIMSS 2003 dataset. Teig et al. (2018) revealed a curvilinear relationship between IBT and science achievement in Norwegian TIMSS 2015 data. In the results, IBT was positively correlated with achievement, but employing highly frequent inquiry activities resulted in decreased achievement.

Beyond the dichotomization of DI and IBT, some studies have emphasized the need for proper balance of the two teaching approaches, which implies that neither instructional method is capable of standing on its own without the incorporation of other methods. Cobern et al. (2010) supported the conclusion that using one method did not show a significant difference in learning outcomes in eighth-grade students, and neither method should be made exclusive in the classroom, and teachers should use the advantages of both methods to increase learning outcomes.

12.2.3 Relationship Between Inquiry Teaching and Science Achievement in PISA Studies

In the 2006 PISA data, there was an inverse relationship between the degree of IBT (described as student investigations and hands-on activities) and science literacy in students in Qatar (Areepattamannil, 2012), Canada (Areepattamannil et al., 2011), and Finland (Lavonen & Laaksonen, 2009). McConney et al. (2014) showed that among 40,000 students across Australia, Canada, and New Zealand, those who experienced a higher frequency of IBT exhibited lower scientific literacy levels. Cairns and Areepattamannil (2019) investigated the relation of inquiry-based science teaching to the science achievement of 170,000 students in 54 countries using Hierarchical Linear Modeling (HLM) analyses. The results revealed that IBT was significantly negatively related to science achievement. Kang and Keinonen (2018) investigated the effects of instructional methods consisting of three types of student-centered approaches—relevant topic based, inquiry based, and discussion based—on science achievement in the Finnish sample. The study obtained mixed results: open inquiry-based learning was indicated as a strong negative predictor of students' performance, whereas guided inquiry-based learning was indicated as a strong positive predictor of students' performance. The authors' interpretation was that not all levels of inquiry are effective for increasing students' knowledge acquisition; thus, teachers should consider proper inquiry strategies based on the purpose of their instruction. For instance, students may often need help with inquiry design because of their lack of procedural knowledge of what and how to investigate in scientific experiments at school. Jiang and McComas's (2015) findings were consistent with those of Kang and Keinonen (2018). Based on the frequencies of four inquiry components (conducting activities, drawing conclusions, designing investigations, and asking questions) shown in the inquiry teaching survey items of PISA 2006, five levels of inquiry teaching were generated (levels 0–4) by the propensity score analysis. At level 0, none of the inquiry components were sufficiently implemented. At level 4, all the components were sufficiently implemented. The study's results revealed that the highest science achievement occurred at level 2 of inquiry teaching, where students frequently conduct activities and draw conclusions. This study concluded that increasing the level of openness in IBT is not beneficial to all students' science learning.

12.3 Methods

12.3.1 Data and Sample

This study used 2015 PISA data. PISA is a global assessment of participating 15-year-old students designed to evaluate their academic performance in three domains: reading, mathematics, and science. The assessment was first administered in 2000 in a 3-year cycle. In each PISA administration, all three domains are assessed, but one domain includes an extensive set of questions. In 2015, the major subject was science. In addition to assessments of reading, mathematics, and science, students and teachers complete a questionnaire on background information, various perceptions, and learning and teaching activities within their schools. PISA thus offers a great opportunity to delve into how different aspects of students, teachers, and schools influence the quality and equity of educational outcomes give the nested nature of the PISA data (OECD, 2017).

This study used three groups of variables: (1) outcome variables regarding the measurement of scientific literacy (i.e., content knowledge and procedural and epistemic knowledge); (2) independent variables regarding teaching practices (IBT and ID); and (3) covariates such as student demographic information, affective factors (e.g., motivation), socioeconomic status, and school characteristics (school type, climate, etc.)

The sixth wave (2015) of PISA assessed 51,934 students from 17,912 schools in 35 OECD countries and 37 partner countries. The sample in the United States used for RQ1 consisted of 5712 students from 177 schools. Of this population, 10th-grade students (about 74%) were the major participants; some were also in grades 7 to 12. The proportion of females (50%) and males (50%) was almost the same.

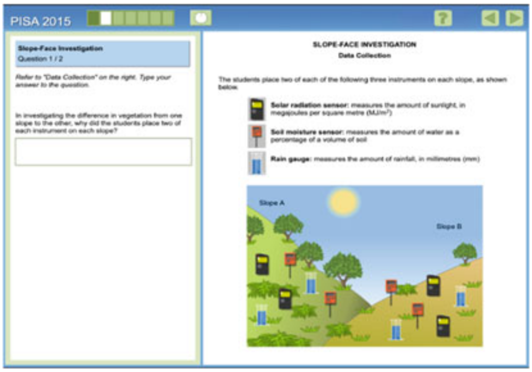
12.3.2 Measures

The dependent variable in the multilevel analysis was 15-year-old-students' scientific literacy. PISA reports test scores for subscales of science literacy: *content knowledge* and *procedural and epistemic knowledge* (hereafter referred to as *P&E knowledge*). The content knowledge items tend to require the application of everyday content knowledge and the ability to recognize aspects of simple scientific phenomena. The P&E knowledge items require sophisticated application to explain hypotheses of novel scientific phenomena, events, and processes (OECD, 2017).

IBT and DI were the main independent variables (Table 12.1) The choice of covariates was based on previous empirical research on school effectiveness and PISA studies (e.g., Anderson et al., 2009; Lam & Lau, 2014; Mostafa, 2010; You, 2015), along with multicollinearity among independent variables for multilevel modeling. The variables included from the student questionnaire were ESCS (index of economic, social, and cultural status), gender, grade, and students' motivation in learning science. The school-related variables were mean school ESCS, school type (public vs. private), school size, science-specific school resources, and school climate (focusing on attitudes, behaviors, and group norms). The number of years teaching and teachers' professional experience from the teacher questionnaire were also included.

In the student questionnaire, question ST098 (When learning school science topics at school, how often do the following activities occur?) consisted of items regarding IBT where students were asked about how frequently specific activities were used in science teaching. Question ST103 included items related to DI. Sample items from the science assessment regarding IBT and DI are shown in Table 12.1.

Table 12.1 Sample items from science assessment and student survey

<table>
<tr><th colspan="6">Instruments and sample items</th></tr>
<tr><td colspan="6"><b><i>A sample item from student science assessment</i></b>


<b>Slope-Face Investigation</b>
Question 1 / 2
<i>Refer to "Data Collection" on the right. Type your answer to the question.</i>
In investigating the difference in vegetation from one slope to the other, why did the students place two of each instrument on each slope?
<b>SLOPE-FACE INVESTIGATION</b>
<b>Data Collection</b>
The students place two of each of the following three instruments on each slope, as shown below.
<b>Solar radiation sensor:</b> measures the amount of sunlight, in megajoules per square metre (MJ/m²)
<b>Soil moisture sensor:</b> measures the amount of water as a percentage of a volume of soil
<b>Rain gauge:</b> measures the amount of rainfall, in millimetres (mm)</td></tr>
<tr><td colspan="6"><b><i>Sample IBT items from student questionnaire</i></b></td></tr>
<tr><td colspan="6"><b>ST098</b> When learning <school science> topics at school, how often do the following activities occur? *1 (<i>In all lessons), 2 (In most lessons),</i> 3 (<i>In some lessons), 4 (Never or hardly ever)</i></td></tr>
<tr><td><b>ST098Q01TA</b></td><td>Students are given opportunities to explain their ideas.</td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td><b>ST098Q02TA</b></td><td>Students spend time in the laboratory doing practical experiments.</td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td><b>ST098Q03TA</b></td><td>Students are required to argue about science questions.</td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td colspan="6"><b><i>Sample DI item from student questionnaire</i></b></td></tr>
<tr><td colspan="6"><b>ST103</b> How often do these things happen in your lessons for this <school science> course?
*<i>1 (Never or almost never), 2 (Some lessons), 3 (Many lessons), 4 (Every lesson or almost every lesson)</i></td></tr>
<tr><td><b>ST103Q01NA</b></td><td>The teacher explains scientific ideas.</td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td><b>ST103Q02NA</b></td><td>A whole class discussion takes place with the teacher.</td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
<tr><td><b>ST103Q03NA</b></td><td>The teacher explains scientific ideas.</td><td>1</td><td>2</td><td>3</td><td>4</td></tr>
</table>

12.3.3 Multivariate Multilevel Model

The PISA survey has a multivariate hierarchical data structure where students are nested into schools and measured by multiple outcomes. With this data frame, one of the assumptions for regression analysis, residual independence, could be violated because the students who attended the same school would be more similar than students who attended different schools. Additionally, because two scientific literacy outcomes, content knowledge, and P&E knowledge, were measured for the same students, there was a possibility that they were related to each other. To consider multilevel and multiple outcomes' dependency simultaneously, the current study used a multivariate multilevel model (Snijders & Bosker, 1999). This model prevents the underestimation of school effects (Borman & Dowling, 2010; Uline & Tschannen-Moran, 2008) and the inflation of the Type I error rate (Baldwin et al., 2014) that can occur when the nested data structure or multiple outcomes' dependency are ignored. Additionally, the multivariate multilevel model can provide variability across students and schools (variance estimates) along with the relations between outcomes (covariance estimates) at each level.

Based on the PISA data structure, a multivariate 3-level (outcomes, students, and schools) model was used, as shown in Fig. 12.2.

Like other empirical data, PISA data also have missing variables, especially at the outcome level. By using a restricted maximum likelihood estimation, we handled the missing data and estimated components of fixed effects and random effects accurately in the multivariate multilevel model (Patterson & Thompson, 1971). The analyses were conducted using the "lme4" package in R software (Bates et al., 2015).

All selected predictors were included in the fully conditional model. Level 1 (the outcome level) is shown as below:

$$Y_{osh} = \pi_{Csh}\, d_{Csh} + \pi_{Psh}\, d_{Psh} \tag{12.1}$$

Here, d_{qsh} was the dummy coded variable corresponding to two outcomes, content (d_{Csh}) and P&E knowledge (d_{Psh}); Y was the two outcomes' observed scores; and π_{osh} was the expected outcome score. That is, when imputing 0 or 1 at d_{osh}, the value of π_{osh} will be the content score or P&E knowledge score. At level 1, we did not have

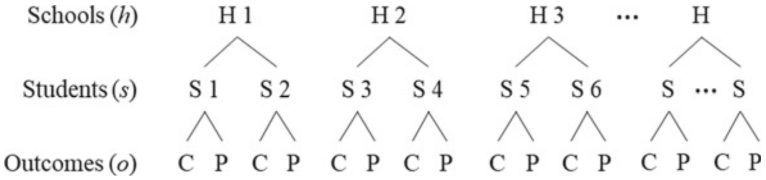


Fig. 12.2 Network depicting outcomes o (level-1) nested in students s (level-2) within schools h (level-3) Note that C represents content knowledge at outcome level and P indicates P&E knowledge score

an error term because level 1 is used as an indicator of the multivariate outcomes' structure in the model (Goldstein, 2010; Hox, 2010).

At level 2 (the student level), the predictors of individual level such as gender, grade, ESCS, and motivation level were included, as shown below:

$$\begin{cases} \pi_{Csh} = \beta_{C0h} + \beta_{C1h}(\text{Gender}) + \beta_{C2h}(\text{Grade}) \ldots + \beta_{Csh}(X) + r_{Csh} \\ \pi_{Psh} = \beta_{P0h} + \beta_{P1h}(\text{Gender}) + \beta_{P2h}(\text{Grade}) \ldots + \beta_{Psd}(X) + r_{Psh} \end{cases}. \tag{12.2}$$

where β_{C0h} and β_{P0h} pooled outcome scores across students within schools, and β_{Csh} and β_{Psh} were coefficients to each of the student-level predictors. The residuals (r_{qsh}) at level 2 were assumed to follow a multivariate normal distribution with means of zero and a variance–covariance matrix specified as follows:

$$\begin{bmatrix} r_{Csh} \\ r_{Psh} \end{bmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{r.C}^2 & \sigma_{r.CP} \\ \sigma_{r.CP} & \sigma_{r.P}^2 \end{pmatrix}\right].$$

Finally, at level 3 (the school level), the predictors of school level such as IBT, direct teaching, and school size were included as shown below:

$$\begin{cases} \beta_{C0h} = \gamma_{C00} + \gamma_{C01}(\text{Inquiry}) + \gamma_{C02}(\text{Direct}) + \gamma_{C03}(\text{Size}) \ldots + \gamma_{C0h}(Z) + u_{C0h} \\ \beta_{P0h} = \gamma_{P00} + \gamma_{P01}(\text{Inquiry}) + \gamma_{P02}(\text{Direct}) + \gamma_{P03}(\text{Size}) \ldots + \gamma_{P0h}(Z) + u_{P0h} \\ \beta_{C1h} = \gamma_{C10}, \\ \beta_{C2h} = \gamma_{C20}, \\ \ldots \\ \beta_{Csh} = \gamma_{Cs0}, \\ \beta_{P1h} = \gamma_{P10}, \\ \beta_{P2h} = \gamma_{P20}, \\ \ldots \\ \beta_{Psh} = \gamma_{Ps0} \end{cases}. \tag{12.3}$$

The set of the two outcomes' level 3 residuals were also assumed to follow a multivariate normal distribution with means of zero and a variance–covariance matrix as follows:

$$\begin{bmatrix} u_{Coh} \\ u_{P0h} \end{bmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u.C}^2 & \sigma_{u.CP} \\ \sigma_{u.CP} & \sigma_{u.P}^2 \end{pmatrix}\right].$$

Note that when we included the predictors, we used grand-mean centering to estimate the effects of the level 3 predictors and the effects of the level 2 predictors separately (Enders & Tofighi, 2007).

12.4 Results

This study aimed to contribute to the current school effect literature by examining the relationship of teaching methods to student academic achievement. For this purpose, student background and classroom, along with school characteristics drawn from the questionnaires, were used as control variables. Furthermore, we compared the implementation of teaching practices in the top-performing five countries and the United States. Our findings could help achieve a comprehensive understanding of which science teaching practices are preferred or not preferred. Such an understanding can provide a basis for determining what matters when engaging students in science education and implementing desired teaching practices.

The standardized average science score of the 5,712 U.S. students from 177 schools in the 2015 PISA was 496 (OECD average: 493) on a scale from 0 to 1000. The mean score of content knowledge was 490.24; the scores ranged from 192 to 813.80. The mean score of P&E knowledge was 500.45; the scores ranged from 206.06 to a maximum of 819.09. The means of two science knowledge domains (content versus P&E) were significantly different ($t(5711) = 31.72$, $p < 0.05$), indicating that the P&E knowledge score (500.45) was higher than the content knowledge score (490.24) (Table 12.2). Additionally, all the continuous variables had acceptable skewness and kurtosis and were assumed to be normally distributed.

Table 12.2 Descriptive statistics for continuous variables

	Min	Max	Mean	SD	Skewness	Kurtosis
Dependent variables						
Content knowledge	192.07	813.80	490.24	96.96	0.09	−0.49
Procedural & epistemic knowledge	206.06	819.09	500.45	93.99	0.02	−0.53
Student level variables						
Motivation	1.00	8.00	5.65	1.29	−0.44	0.28
ESCS	−3.79	2.97	0.08	1.00	−0.44	−0.20
School level variables						
Teaching experience (yrs)	3.00	23.79	13.95	3.54	−0.35	0.52
PD participation	2.50	4.67	3.60	0.38	−0.19	−0.02
Size	22.00	4230.00	1367.93	874.72	0.72	0.23
Resources	0.00	8.00	5.38	1.86	−0.88	0.57
Climate	1.00	4.00	2.93	0.46	−0.25	1.48
Mean ESCS	−1.65	1.13	0.08	0.54	−0.56	0.59
Inquiry based teaching	1.00	4.00	2.47	0.70	0.39	−0.35
Direct instruction	1.00	4.00	2.80	0.81	−0.24	−0.72

Table 12.3 provides the frequency of categorical variables at each level and the mean scores of both outcomes (content and P&E). For all categorical variables, significance tests were conducted to examine the differences between outcome scores by categories. According to the results, both content ($F(4, 5076) = 98.55$), $p < 0.001$) and P&E ($F(4, 5076) = 105.84, p < .001$) scores were significantly different across grade levels. There was a significant gender difference for content knowledge ($t(5710) = 6.80, p < 0.05$) but a nonsignificant difference ($t(5710) = 0.33, p > 0.05$) for P&E knowledge. Additionally, there were significant differences by school type for content ($t(5710) = 3.32$, $p < 0.05$) and P&E knowledge ($t(5710) = 3.58$, $p < 0.05$). Last, there were significant differences by race and ethnicityfor content ($F(3, 5686) = 342.43$, $p < 0.001$) and P&E knowledge ($F(3, 5686) = 287.17, p < .001$).

Before testing the full conditional model (Eqs. 12.1, 12.2, and 12.3), the null model (not including any predictors) was analyzed to estimate the intraclass correlation (ICC) and variability across students and schools. The ICC for the null model was 21.4% (ICC = 0.214) for content knowledge and 20.9% (ICC = 0.209) for P&E

Table 12.3 Descriptive statistics for categorical variables and mean and SD of dependent variables by categories

	Frequency		Content		P&E	
	N	%	Mean	SD	Mean	SD
Student level variables						
Grade						
7th	1	0.0	355.15	71.95	399.80	76.25
8th	12	0.2	381.86	81.12	403.41	79.46
9th	529	9.3	416.05	95.28	426.23	92.28
10th	4210	73.7	496.06	93.57	505.82	89.20
11th	953	16.7	507.13	146.12	519.20	151.82
12th	7	0.1	502.27	71.95	504.59	76.25
Gender						
Female	2854	50.0	481.54	92.86	500.04	90.54
Male	2858	50.0	498.93	100.15	500.86	97.32
Race/Ethnicity						
White	2498	43.7	528.07	90.43	533.51	88.41
Hispanic	1761	13.8	426.88	77.76	439.43	77.66
Black	790	30.8	462.29	89.20	478.26	86.98
Other	641	11.2	499.29	97.22	509.82	96.26
School level variables						
School type						
Private	310	5.5	508.43	86.30	519.40	82.70
Public	5303	94.5	489.67	97.39	499.80	94.46

knowledge. Generally, the fact that the variance in schools was over 10% indicates that the multilevel model should be used (Lee, 2000).

Between-student variabilities for content knowledge ($\sigma^2_{r.C}$ = 7419.59, SD = 86.17) and P&E knowledge ($\sigma^2_{r.P}$ = 7008.64, SD = 83.72) were statistically significant in the model. Between-school variance estimates were also statistically significant for content knowledge ($\sigma^2_{u.C}$ = 2021.23, SD = 45.11) and P&E knowledge ($\sigma^2_{u.P}$ = 1846.53, SD = 43.11). This indicated that students can have very different scores for content and P&E knowledge even within the same school. Similarly, schools' mean scores for content and P&E knowledge are also substantially varied. To explore the effects of IBT and DI on scientific literacy, we had to include, as controls in the full model, student-level and school-level predictors that explained the variability. To control the school and student variables that generated this variability and the factors to explain the variability, we included student-level and school-level predictors in the full model.

According to the results of the full conditional multilevel model (Table 12.4), when controlling for student-level variables (grade levels, gender, motivation, ESCS, race/ethnicity) and school-level variables (teaching experience, PD participation, school size, school type, resources, climate, school mean ESCS), IBT ($b = -19.48$, $p < 0.05$ for content; $b = -20.70$, $p < 0.05$ for P&E) and both content and P&E knowledge showed an inverse relationship. Further, DI (b = 16.12, $p < 0.05$ for content; b = 19.51, $p < 0.05$ for P&E) was positively associated with the two outcomes. All the student-level covariates had significant impacts on content and P&E knowledge scores. For both outcome scores, for example, higher grade (b = 31.75, $p < 0.05$ for content; b = 31.19, $p < 0.05$ for P&E), male (b = 22.68, $p < 0.05$ for content; b = 6.62, $p < 0.05$ for P&E), higher level of motivation (b = 6.43, $p < 0.05$ for content; b = 6.26, $p < 0.05$ for P&E), higher level of ESCS (b = 19.45, $p < 0.05$ for content; b = 18.72, $p < 0.05$ for P&E), and white students (b = 18.90, $p < 0.05$ for content; b = 14.13, $p < 0.05$ for P&E) were related to higher content and P&E knowledge scores. At the same time, Hispanic ($b = -52.91$, $p < 0.05$ for content; $b = -53.25$, $p < 0.05$ for P&E) and Black students ($b = -11.50$, $p < 0.05$ for content; $b = -9.84$, $p < 0.05$ for P&E) were related to lower outcome scores. For school-level covariates, public school (b = 26.97, $p < 0.05$ for content; b = 23.69, $p < 0.05$ for P&E) and higher mean ESCS (b = 24.30, $p < 0.05$ for content; b = 23.63, $p < 0.05$ for P&E) were positively associated with content and P&E knowledge scores. Last, PD participation (b = 12.06, $p < 0.05$) and climate (b = 11.40, $p < 0.05$) were only positively associated only with the P&E knowledge score.

12.5 Discussion

Many scholars are concerned that American students show no improvement in science on international exams. There is a substantial amount of research on the ways to improve science achievement. This study provides implications and recommendations to teachers and educators that focus on the effectiveness of the two

Table 12.4 Results of the multivariate multilevel model

	Content		P&E		
	Est.	*(S.E.)*	Est.	*(S.E.)*	η^2
Fixed Effects (Intercept)	458.30*	9.52	481.08*	9.35	0.984
Student-Level					
Grade (Grade 7 = 0)	31.75*	2.42	31.19*	2.36	0.153
Gender (Female = 0)	22.68*	2.33	6.62*	2.27	0.402
Motivation	6.43*	1.59	6.26*	1.56	0.015
ESCS	19.45*	1.42	18.72*	1.38	0.276
Ethnicity (White = 1)	18.90*	4.09	14.13*	4.00	0.176
Ethnicity (Hispanic=1)	−52.91*	5.11	−53.25*	5.00	0.101
Ethnicity (Black = 1)	−11.50*	4.53	−9.84*	4.43	0.011
School-Level					
Teaching experience	1.12	0.62	0.80	0.61	0.010
PD participation	11.02	5.24	12.06*	5.14	0.007
Size	0.00	0.00	0.00	0.00	0.003
Resources	0.28	1.09	0.02	1.07	0.002
Type (Private = 0)	26.97*	9.11	23.69*	8.95	0.001
Climate	9.46	4.64	11.40*	4.56	0.011
Mean ESCS	24.30*	4.44	23.63*	4.36	0.029
Inquiry-based teaching	−19.48*	1.84	−20.70*	1.80	0.061
Direct instruction	16.12*	1.53	15.91*	1.50	0.084
Random Effects **Student-Level**					
Variance $(\sigma^2_{r.C}, \sigma^2_{r.C})$	5984.45		5713.39		
Covariance $(\sigma_{r.CP})$		4853.35			
School-Level					
Variance $(\sigma^2_{u.C}, \sigma^2_{u.P})$	353.46		343.22		
Covariance $(\sigma_{u.CP})$		289.08			

Note Est. = Estimates, * $p < 0.05$

specific teaching methods, IBT and DI. Extensive research has presented evidence of the importance and effectiveness of IBT over many years. In addition to "knowing" science concepts, students should be expected to use their understanding to investigate the natural world through the practices of scientific inquiry and solve meaningful problems using engineering design practices. As a result, IBT has been supported by many research projects and U.S. national documents (e.g., NSES, the Framework, and NGSS), but some recent empirical evidence has shown different views of IBT. In recent PISA studies, a negative relationship between IBT and science achievement in the contexts of different countries has often been observed (e.g., Cairns & Areepattamannil, 2019; Liou, 2020). The current study also obtained similar findings, including the finding that inquiry practice does not have its most profound effect on science achievement; rather, DI may support students' scientific literacy more successfully. Moreover, the findings revealed that U.S. teachers use more diverse inquiry-oriented teaching methods compared to the five top-performing countries and regions. Considering the many advantages of IBT and its current usage, the United States can and should do a much better job in terms of student achievement. However among the 35 industrialized nations that are members of the OECD, the U.S. ranked 19th in 2006. Although we do not argue that IBT does not have its place for learning of knowledge and skills, our findings do tip the balance toward the DI end of the scale for science learning. To make sense of why DI is associated with more productive outcomes and IBT has a negative relationship with outcomes, we need to contemplate the key issues and empirical evidence at stake in debating DI vs. IBT. What are the disadvantages of the IBT approach for students' science learning? How can they be addressed? In contrast, what are the strengths of DI for science learning? How can they be incorporated into science courses for productive outcomes? Inquiry is regarded as a self-directed form of learning (Bencze & Di Giuseppe, 2006). If a student is not comfortable with taking responsibility for their own learning, there would be no takeaways from this form of instruction. Kirschner et al. (2006) pointed out that IBT can work well only to improve students' affective aspects such as interests and motivation. Enjoyment and learning without guidance increase the cognitive load of students, possibly preventing students from grasping the main concepts being taught.

PISA is a test of application and not of simple recall or inquiry-type skills; it measures students' scientific knowledge (knowledge of science and knowledge about science) and the use of that knowledge to explain scientific phenomena and draw evidence-based conclusions about scientific issues (OECD, 2019). To yield excellent results in student performance in a test such as the PISA, DI is the best teaching method. It helps prepare students for standardized or other formal tests, helping them gain and retain a wealth of knowledge and master concepts in limited class time (Liou, 2020). However, a report by McKinsey and Company (Mourshed et al., 2017) suggested a different view, digging deeper into PISA data. The report argued that the best performance is obtained when the two teaching styles work together. According to the report, the "sweet spot" (pp. 7–8) combines DI in most-to-all science classes and IBT in some-to-many classes.

12.5.1 Implications for Teacher Education and Professional Development

The findings of this study point to the usefulness of the DI method, but we do not want to drift from a rational position to a more one-sided view. Rather than abandoning one teaching approach in science classrooms, this study encourages U.S. science teachers to understand the education systems of other countries (especially top-performing countries in the PISA assessment) and their instructional practices, and to contemplate which education model is most useful to adapt to the U.S. education system. Teachers' instructional practices are naturally affected by their own educational system, history, and culture. For example, most top-performing countries have an educational system that places a high importance on testing. The countries' testing culture is most intense when teachers are inclined to choose DI. Thus, it is inevitable that DI is prevalent in science classrooms in top-performing countries. In contrast, U.S. science education values inquiry-based reasoning and meaning making rather than focusing on performing well on international tests and high-stakes exams. Thus, adapting or emulating teaching practices of other countries with different educational philosophies and cultures cannot be a good way to achieve the goal for the improvement of scientific literacy.

This study concludes that maximizing the advantages of the two methods by combining them is the best way to achieve effective science teaching in the United States. Losardo and Bricker (1994) found that DI works best when accompanied by IBT. Generally, DI does not preclude teaching students how to find problems, solve problems, think about science in critical ways, collaborate with others, or take charge of their own learning. Houseal et al. (2016) also supported the argument that teachers must use a combination of DI and inquiry-based learning to help students form an understanding of the scientific world.

Even though teachers recognize the importance and necessity of IBT, there may be a gap between what they know and desire and what they actually do. Thus, continuing professional development provides teachers with a clear vision of what best teaching practices look like. Because students at different achievement levels need different kinds of inquiry-learning activities, the combined approach can strengthen students' scientific literacy. Although low performers need more guidance on IBT, high performers may need little-to-no support from their teachers in designing and conducting their investigations from scratch. Thus, if utilized appropriately and in a purposeful manner that considers students' prior knowledge and lesson content, IBT can produce positive learning effects in the science classroom. Kirschner et al. (2006) asserted that for novice learners, minimally guided instruction is likely to be ineffective. Thus, scaffolding and guidance need to be emphasized when inquiry-based learning strategies are used because their effects on achievement can be considerable.

12.5.2 Limitations and Future Directions

The PISA dataset provides the largest international sample with more than 60 participating countries, and together these countries represent nearly 90% of the world economy. Data that are internationally available can be compared and contrasted, and the nested feature of data (i.e., students nested in a school) can help generate more reliable results. However, the PISA dataset has some inherent limitations. The PISA measure of teaching practices is based on self-reports and asked about the frequency of teaching practices. This makes us expect a considerable variation in the success of IBT or DI. Additionally, this study used sophisticated procedures to analyze data and examine its hypotheses, but PISA data are cross-sectional in nature. Therefore, the directions of any possible changes over a period of time cannot be ascertained with the data. Next, because the data do not reflect students' prior performance, which can be controlled in a statistical model, this can prevent the formulation of more exact predictions. To confirm this study's findings and understand the hidden processes that shape students' cognitive advancement, future researchers can conduct a longitudinal examination of the relationship between IBT or DI and science achievement. Further, the analysis of secondary data was limited to existing variables in this study. Third, important variables were not available for the analysis. Another possibility for future studies is to investigate other significant student-, classroom- and school-level variables to improve the current model. This would provide more insightful information about compound educational environments.

12.6 Conclusion

Like previous studies, this study provided clear evidence that employing DI results in significant gains in science achievement. It also found a significant negative relationship between IBT and science achievement. These findings should motivate teachers and policymakers to understand the necessity and importance of DI for obtaining high scores in international tests. Students should first know and understand basic and fundamental knowledge before moving on to inquiry processes; as the NRC (2012) puts it, "Science is not just a body of knowledge that reflects current understanding of the world; it is also a set of practices used to establish, extend, and refine that knowledge" (p. 26). It seems that DI is more useful for teaching knowledge in an efficient way, especially in lower grades or for low performers. Thus, we conclude that IBT remains a key element of science teaching even though all students need different kinds and levels of inquiry. However, it would be misleading to say that all science should be taught using IBT to help students construct their own knowledge. This finding and the current status of science teachers' implementation of IBT and DI have the potential to stimulate future research.

References

Adak, S. (2017). Effectiveness of constructivist approach on academic achievement in science at secondary level. *Educational Research and Reviews, 12*(22), 1074–1079.

Anderson, J. O., Milford, T., & Ross, S. P. (2009). Multilevel modeling with HLM: Taking a second look at PISA. In *Quality research in literacy and science education* (pp. 263–286). Springer.

Areepattamannil, S. (2012). Effects of inquiry-based science instruction on science achievement and interest in science: Evidence from Qatar. *The Journal of Educational Research, 105*(2), 134–146.

Areepattamannil, S., Freeman, J. G., & Klinger, D. A. (2011). Influence of motivation, self-beliefs, and instructional practices on science achievement of adolescents in Canada. *Social Psychology of Education, 14*(2), 233–259.

Baldwin, S. A., Imel, Z. E., Braithwaite, S. R., & Atkins, D. C. (2014). Analyzing multiple outcomes in clinical research using multivariate multilevel models. *Journal of Consulting and Clinical Psychology, 82*(5), 920–930.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-8. http://CRAN.project.org/package=lme4

Bencze, J. L., & Di Giuseppe, M. (2006). Explorations of a paradox in curriculum control: Resistance to open-ended science inquiry in a school for self-directed learning. *Interchange, 37*(4), 333–361.

Borman, G., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record, 112*(5), 1201–1246.

Böttcher, F., & Meisert, A. (2013). Effects of direct and indirect instruction on fostering decision-making competence in socio scientific issues. *Research in Science Education, 43*(2), 479–506.

Cairns, D., & Areepattamannil, S. (2019). Exploring the relations of inquiry-based teaching to science achievement and dispositions in 54 countries. *Research in Science Education, 49*(1), 1–23.

Cobern, W. W., Schuster, D., Adams, B., Applegate, B., Skjold, B., Undreiu, A., ... & Gobert, J. D. (2010). Experimental comparison of inquiry and direct instruction in science. *Research in Science & Technological Education*, *28*(1), 81–96.

Dean, D., Jr., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education, 91*(3), 384–397.

Edelson, D. C., Gordin, D. N., & Pea, R. D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the Learning Sciences, 8*(3–4), 391–450.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138.

Esler, W. K., & Sciortino, P. (1991). *Methods for teaching: An overview of current practices.* Contemporary Publishing Company.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research, 82*(3), 300–329.

Goldstein, H. (2010). *Multilevel statistical models* (4th ed.). Hodder Arnold.

Houseal, A., Gillis, V., Helmsing, M., & Hutchison, L. (2016). Disciplinary literacy through the lens of the next generation science standards. *Journal of Adolescent &amp; Adult Literacy, 59*(4), 377–384.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Taylor & Francis.

Jiang, F., & McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and attitudes: Evidence from propensity score analysis of PISA data. *International Journal of Science Education, 37*(3), 554–576.

Kang, J., & Keinonen, T. (2018). The effect of student-centered approaches on students' interest and achievement in science: Relevant topic-based, open and guide dinquiry-based, and discussion-based approaches. *Research in Science Education, 48*(4), 865–885.

Kaya, S., & Rice, D. C. (2010). Multilevel effects of student and classroom factors on elementary science achievement in five countries. *International Journal of Science Education, 32*(10), 1337–1363.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*(10), 661–667.

Lam, T. Y. P., & Lau, K. C. (2014). Examining factors affecting science achievement of Hong Kong in PISA 2006 using hierarchical linear modeling. *International Journal of Science Education, 36*(15), 2463–2480.

Lavonen, J., & Laaksonen, S. (2009). Context of teaching and learning school science in Finland: Reflections on PISA 2006 results. *Journal of Research in Science Teaching, 46*(8), 922–944.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*(2), 125–141.

Liou, P. Y. (2020). Students' attitudes toward science and science achievement: An analysis of the differential effects of science instructional practices. *Journal of Research in Science Teaching, 58*(3), 310–334.

Losardo, A., & Bricker, D. (1994). Activity-based intervention and direct instruction: A comparison study. *American Journal on Mental Retardation, 98*, 744–765.

McConney, A., Oliver, M. C., Woods-McConney, A. M. A. N. D. A., Schibeci, R., & Maor, D. (2014). Inquiry, engagement, and literacy in science: A retrospective, cross-national analysis using PISA 2006. *Science Education, 98*(6), 963–980.

Medrich, E. A., & Griffith, J. E. (1992). *International mathematics and sciences assessments: What have we learned?* Office of Educational Research and Improvement and National Center for Education Statistics, U.S. Department of Education (Report No. NCES92-011, 1992).

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—What is it and does it matter? Results from research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474–496.

Mourshed, M., Krawitz, M., & Dorn, E. (2017). *How to improve student educational outcomes: New insights from data analytics.* McKinsey & Company. https://www.mckinsey.com/~/media/McKinsey/Industries/Social%20Sector/Our%20Insights/How%20to%20improve%20student%20educational%20outcomes/How-to-improve-student-educational-outcomes-New-insights-from-data-analytics.pdf

Mostafa, T. (2010). Decomposing inequalities in performance scores: The role of student background, peer effects and school characteristics. *International Review of Education, 56*(5–6), 567–589.

National Research Council. (1996). *National science education standards.* National Academy Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, cross cutting concepts, and core ideas.* National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* National Academies Press.

OECD. (2017). *PISA 2015 technical report*. http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

OECD. (2019). *PISA 2018 results (Volume I): What students know and can do PISA*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58*(3), 545–554.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T.-Y., & Lee, Y.-H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44*(10), 1436–1460.

Schuster, D., Cobern, W. W., Adams, B. A., Undreiu, A., & Pleasants, B. (2018). Learning of core disciplinary ideas: Efficacy comparison of two contrasting modes of science instruction. *Research in Science Education, 48*(2), 389–435.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Stender, A., Schwichow, M., Zimmerman, C., & Härtig, H. (2018). Making inquiry-based science learning visible: The influence of CVS and cognitive skills on content knowledge learning in guided inquiry. *International Journal of Science Education, 40*(15), 1812–1831.

Teig, N., Scherer, R., & Nilsen, T. (2018). More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science. *Learning and Instruction, 56*, 20–29.

Uline, C., & Tschannen-Moran, M. (2008). The walls speak: The interplay of quality facilities, school climate, and student achievement. *Journal of Educational Administration, 46*(1), 55–73.

Wolf, S. J., & Fraser, B. J. (2008). Learning environment, attitudes and achievement among middle-school science students using inquiry-based laboratory activities. *Research in Science Education, 38*(3), 321–341.

You, H. S. (2015). Do schools make a difference?: Exploring school effects on mathematics achievement in PISA 2012 using hierarchical linear modeling. *Journal of Educational Evaluation, 28*(5), 1301–1327.

Hyesun You earned her B.S. in Chemistry and M.S. in science education from Yonsei University. Her M.Ed in quantitative methods and Ph.D. in Science Education at the University of Texas at Austin. She has since worked as a post-doctoral fellow at New York University and Michigan State University, where she participated in NSF-funded grant projects. She also worked as an assistant professor at Arkansas Tech University by 2021 and she is currently serving as an assistant professor at the University of Iowa. Her research interests center upon interdisciplinary learning and teaching and technology-integrated teaching practices in STEM education. In her dissertation work, she developed and validated a new interdisciplinary assessment in the context of carbon cycling for high school and college students using Item Response Theory. She is also interested in developing technology-embedded curricula and teaching practices in a reform-oriented approach.

Chapter 13
Family Meals and Academic Performance: A Multilevel Analysis for Spain

Nerea Gómez-Fernández and Juan-Francisco Albert

Abstract Students' eating habits have been frequently studied in previous literature as determinants of children and adolescents' academic performance from a health and nutrition point of view. The objective of this chapter is to analyze additional benefits related to student meals by understanding the importance of the family mealtime. Specifically, the aim is to analyze whether the frequency of shared family meals is related to the academic performance of adolescents. To do so, we analyze the data for Spain in PISA 2015. In order to perform a rigorous analysis of the data, we estimate multilevel models that consider the hierarchical PISA data structure: (1) first, public and private schools are randomly selected; and (2) then fifteen-year-old students from the selected schools are selected. The results show that there is a positive relationship between the frequency with which parents eat the main meal with their children and academic performance in reading comprehension as measured by PISA test scores. The positive association is of similar magnitude irrespective of gender and socio-economic and cultural status of the student.

Keywords Mealtime · Family · PISA · Multilevel

13.1 Introduction

The application of multilevel techniques is highly recommended in the analysis of educational data, as it allows us to consider the effect that groups or context can have on individual student outcomes. In the case at hand, the aim of this chapter is to apply multilevel techniques to the analysis of educational data from the international PISA 2015 database. The ultimate goal of the estimations in this chapter is to answer the

N. Gómez-Fernández (✉)
Universitat Politècnica de València, Valencia, Spain
e-mail: negofer@upv.es

Centro Universitario EDEM-Escuela de Empresarios, Valencia, Spain

J.-F. Albert
Universitat de València, Valencia, Spain

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_13

question: Is the frequency of shared family meals related to the academic performance of adolescents?

In recent years, several studies have analysed the relationship between the frequency of family meals and children's and adolescents' behavior in different areas (Lee et al., 2020; Middleton et al., 2020; Sen, 2010), suggesting mostly positive effects of a higher frequency of family meals. For example, family meals have been shown to improve family closeness and the emotional well-being of all family members (Satter, 1986). Family connectedness also benefits from the sense of belonging that develops when children and young people have a family mealtime routine (Fiese et al., 2002). In addition, family meals also have positive effects on parental well-being. Specifically, previous research suggests that parents who had jobs that did not interfere with family mealtimes reported a stronger relationship with their children and spouse and that job satisfaction is linked to having time to come home and join in family meals (Jacob et al., 2008). Family meals have also been shown to have a positive impact on nutrition, with children and young people who eat frequent meals with their families consuming more vegetables, fruit, cereals, calcium- and macronutrient-rich foods, and less soft drinks (Burgess-Champoux et al., 2009; Neumark-Sztainer et al., 2010). In the same vein, studies have shown that the frequency of family meals may also be associated with a lower risk of obesity (Anderson & Whitaker, 2010; Fulkerson et al., 2009). Family meals are also associated with a prevention of high-risk behaviors. In particular, several studies show that children and adolescents who enjoy frequent family meals are less likely to smoke, drink, or use drugs (Eisenberg et al., 2004; Fulkerson et al., 2006; Neumark-Sztainer et al., 2010; Sen, 2010).

The aim of this research is to provide empirical evidence on the relationship among family meals and academic performance. While there are some previous studies that have analysed this relationship (Cullen & Baranowski, 2000; Eisenberg et al., 2004; Kim & Lee, 2021; Miller et al., 2012; Neumark-Sztainer et al., 2010; Shin et al., 2017), our research presents important novelties: (1) it is the first research that analyses this issue using data from the Program for International Student Assessment (PISA), one of the most relevant assessments at the international level; (2) it is the first study conducted for the specific case of Spain; (3) the application of multilevel techniques is novel and overcomes many of the limitations of previous research; (4) the study is conducted distinguishing by competences (mathematics, reading, science, financial literacy, and collaborative problem solving); and (3) an additional analysis is added to investigate whether the relationship between the frequency of family meals and academic performance differs according to the gender or socioeconomic and cultural status of the student.

The chapter proceeds in the following way. We begin with a section aimed at reviewing the most relevant studies in the area in order to know the state of the art. A description of the data used is presented below, as well as a detailed description of the methodology used in the estimation of the models. We then provide and interpret our results and finally we make a series of recommendations based on the results obtained.

13.2 Literature Review

There are few studies that have focused on the effects of family meals on academic performance. The first of the studies in this area was the work of Cullen and Baranowski (2000). These authors analysed 120 boys and girls ages 7–11 in the United States and found that the children who excelled in school more frequently came from homes that partook in frequent family meals. More recently, we find mainly five relevant studies: Eisenberg et al. (2004); Neumark-Sztainer et al. (2010); Miller et al. (2012); Shin et al. (2017); and Kim and Lee (2021). The following is a brief explanation of the research carried out by these authors in order to provide an overview of the state of the art on the subject of this research.

Eisenberg et al. (2004) analysed the association between the frequency of family meals and multiple indicators of adolescent health and well-being using data from a 1998–1999 school-based survey of 4,746 adolescents in the Minneapolis/St Paul metropolitan area. The authors used logistic regressions and found that frequency of family meals was inversely associated with low grade point average. The study by Neumark-Sztainer et al. (2010) analysed the effects of the EAT project (Eating Among Teens), a large population based study of adolescents in Minnesota. Specifically, the Project EAT-I survey was completed by 4,746 middle-school and high-school students, and the Project EAT-II longitudinal survey, was completed by 2,516 of the original participants five years later. The authors used descriptive statistics and suggested that family meals tend to be positively associated with grades. However, they highlight the need for further research in order to elucidate the pathways that underpin the relationships between family meals and academic performance. In this vein, Miller et al. (2012) analysed data from a panel sample of 21,400 children aged 5–15 in the United States. The authors conducted a rigorous analysis by examining individual students in separate age groups in order to establish whether the results differ according to the age of the children. Applying fixed effects models, the authors found that there are no statistically significant relationships between the frequency of family meals and academic performance. This absence of a relationship is a novel result that contradicts the findings of other research in the area. The results did not vary according to the age of the child. More recently, Shin et al. (2017) analyzed the data of 302 participants that were recruited from a middle school at Goyangsi (South Korea) and applied multiple regression techniques. The authors found that engagement in family meals was related to better eating behavior, academic achievement, and quality of life among middle-school students. However, the authors highlight the need for further studies to support the benefit of family meals in improving academic achievement among high-school students as well as middle-school students. Very recently, Kim and Lee (2021) analysed 241 data collected through self-administered questionnaires for middle-school students in Daegu (Korea). Using descriptive statistics, the authors found that the family meal frequency was significantly and positively related to middle-school students' academic outcomes.

The review of previous literature in this area shows that further research on the effects of family meals on academic performance is needed for various reasons.

Firstly, because considering cultural differences, it is relevant to extend the analysis to different geographical areas from those already analysed (United States and Korea) to see whether the conclusions reached in culturally different contexts are the same or different. On the other hand, methodologically speaking, none of the statistical analyses of the above-mentioned research take into account the multilevel structure of the educational data when performing the corresponding analyses. Finally, the scarcity of empirical evidence per se and, in particular, the lack of recent studies in the area, highlight the need for further research. In addition, as discussed in the introduction, this research brings important novelties, including the consideration and comparison of different competences and a complementary analysis to investigate whether the relationship between the frequency of family meals and academic performance differs according to the gender or socio-economic and cultural status of the student.

13.3 Methodological Approach

13.3.1 Variables

The aim of this research is to analyze whether the frequency of shared family meals is related to the academic performance of 15-year-old students. Therefore, the dependent variable of our models to be estimated is the academic performance which is measured on this occasion through the score achieved in the mathematics, reading comprehension, science, financial literacy, and collaborative problem solving tests of the PISA 2015 survey.

In order to provide the reader with a clearer understanding of what each of these competences reflects, a brief explanation of what lies behind each of the competences assessed is given below (OECD, 2015): (1) Mathematical competence assesses the student's capacity to formulate, employ and interpret mathematics in a variety of contexts; (2) competence in reading comprehension assesses the student's capacity to understand, use, reflect on and engage with written texts in order to achieve one's goals, develop one's knowledge and potential and to participate in society; (3) competence in science reflects the ability to engage with science-related issues and with the ideas of science as a reflective citizen; (4) assessment of financial literacy draws on a range of knowledge and skills related to the development of the capacity to deal with the financial demands of everyday life and uncertain futures within contemporary society; and (5) the collaborative problem solving assessment measures students' capacity to successfully engage in a process whereby two or more agents attempt to solve a problem by sharing the understanding and effort required to come to a solution, and pooling their knowledge, skills, and efforts to reach that solution.

In the PISA test, each student takes a different combination of test items. Therefore, in order to establish a common and comparable scale of performance measurement across all students, final scores are estimated as plausible values using Item

Response Theory (OECD, 2015). In this research, we follow Säälik et al. (2015) and Rutkowski et al. (2019) and we use as dependent variables the first plausible values: *PV1MATH; PV1READ*; *PV1SCIE*; *PV1FLIT*; and *PV1CLPS*. Scores are scaled so that the OECD average in each domain is 500 and the standard deviation is 100. The decision to use only the first plausible value is based on research by Jerrim et al. (2017). These authors conclude that the use of one plausible value or all plausible values has no impact upon the results. Likewise, the PISA survey organizers themselves (OECD, 2009b, p. 46) recognize that the use of a single plausible value provides unbiased point and sampling variance estimates on large samples, as is the case in this research.

On the other hand, the explanatory independent variable is the variable *PA003Q02TA* from the PISA questionnaire that reflects parents' responses to the question: Activities with your child, how often: Eat <the main meal> with my child around a table. We use data from PISA 2015 as this is the most recent database in which information is available for the variable *PA003Q02TA* in Spain since in the most recent data from 2018, Spanish parents did not answer this question. Regarding the response options, the original variable includes 5 possible answers: (1) never or hardly ever; (2) once or twice a year; (3) once or twice a month; (4) once or twice a week; and (5) every day or almost every day. Analyzing the distribution of responses to this question (Table 13.1), we have chosen to group the first three response options, so that the variable included in our regressions takes values from 0 to 2. A value of 0 indicates that parents never/hardly ever/once or twice a year/once or twice a month eat with their children. A value of 1 indicates that parents eat with their children once or twice a week, and a value of 2 indicates that they eat together every day or almost every day.

In addition, considering previous literature, we have included a series of control variables related to socio-demographic characteristics of the students and the school they attend: the student's gender (dummy variable for *female* gender); information on whether the student has ever repeated a year (*repeat*); immigration status index (*immig*); index of economic, social, and cultural status (*ESCS*); age of beginning of ISCED0 (*ISCED0*); school ownership (*schltype*); and class size (*clsize*).

A brief justification for the inclusion of the previously mentioned control variables is provided below. At the student level, we have included a gender variable as previous literature has shown that there is a gender gap in academic performance between

Table 13.1 Original distribution of responses for variable PA003Q02TA

Activities with your child, how often: Eat <the main meal> with my child around a table	Frequency	Percent
Never or hardly ever	23	0.49
Once or twice a year	11	0.23
Once or twice a month	29	0.62
Once or twice a week	286	6.09
Every day or almost every day	4,346	92.57

males and females (Parker et al., 2018). Controlling for grade repetition is also important, as previous research has shown that repeaters tend to perform worse academically than non-repeaters (Ikeda & García, 2014). Regarding the immigration status index, in the specific case of Spain, several studies have suggested that immigrant students perform worse academically than native students, even after controlling for the socio-economic status and language skills. At the family level, we consider the "Index of economic, social and cultural status" (ESCS) since previous literature has shown that socio-economic status is one of the key variables in explaining the academic performance of schoolchildren (Suárez-Álvarez et al., 2014; White, 1982). In Pisa 2015, ESCS is a composite score built using Item Response Theory scaling and the indicators: parental education, highest parental occupation, and home possessions including books in the home. The starting age of ISCED0 has also been considered as previous research has shown that students who voluntarily enter the education system earlier achieve higher levels of academic performance at compulsory levels of education (Robbin, 1996; Kashkary, 2012).

At the school level, we have controlled for school ownership (public, private government dependent, or private independent) as previous studies in Spain have shown that on average students in public schools perform worse than those in private schools (Choi & Calero, 2012). We have also controlled for the number of students in the classroom, since a negative relationship between the number of students per teacher and academic performance has been shown (Koc & Celik, 2015). The main descriptive statistics for all variables included in the regressions are given in Table 13.2.

13.3.2 Multilevel Modeling

In order to perform a rigorous analysis of the data, we estimate multilevel models that consider the hierarchical PISA data structure: (1) first, public and private schools are randomly selected; and (2) then fifteen-year-old students from the selected schools are selected. This multilevel structure implies that that the independence principle is not met since there is dependence on observations within each school (Hox, 1995). As a consequence, it is not appropriate to use the Ordinary least squares (OLS) regression since we would be incurring in the atomistic fallacy (Alker, 1969) and forgetting that the school-level context of the students matters and that there is a significant degree of homogeneity among students attending the same school. Considering this hierarchical structure, previous literature has pointed to the desirability of employing multilevel regression techniques to examine PISA data (Gómez-Fernández & Mediavilla, 2021; OECD, 2009a; Thorpe, 2006).

In this chapter, we follow Snijders (2011) and Gómez-Fernández and Mediavilla (2021) and estimate the multilevel model presented in Eqs. (13.1) and (13.2) for the five competences evaluated in this research: mathematics, reading, science, financial literacy, and collaborative problem solving.

Table 13.2 Descriptive statistics for dependent, independent, and control variables

Variable	Obs	Mean/%	Std. Dev.	Min	Max
Dependent variables					
PV1MATH	6,736	490.438	82.914	182.21	763.9
PV1READ	6,736	499.495	85.522	161.767	779.974
PV1SCIE	6,736	496.987	86.400	210.696	754.33
PV1FLIT	6,736	472.335	100.837	0	788.007
PV1CLPS	6,736	500.756	87.858	142.154	800.918
Explanatory independent variable					
PA003Q02TA	4,695	1.910	0.327	0	2
0: From never to once or twice a month	63	1.34%			
1: Once or twice a week	286	6.09%			
2: Every day or almost every day	4,346	92.57%			
Control independent variables					
Female (dummy)	6,736	0.505	0.500	0	1
Repeat (dummy)	6,699	0.266	0.442	0	1
Immig	6,577	0.188	0.567	0	2
0: Native	5,896	89.65%			
1: Second-Generation	125	1.90%			
2: First-Generation	556	8.45%			
ESCS	6,678	−0.449	1.186	−4.352	3.091
ISCED0	6,301	1.799	0.934	0	6
0: 1 year or younger	518	8.22%			
1: 2 years	1,249	19.82%			
2: 3 years	3,965	62.93%			
3: 4 years	311	4.94%			
4: 5 years	128	2.03%			
5: 6 years or older	68	1.08%			
6: I did not attend <ISCED 0>	62	0.98%			
Schltype	6,575	1.604	0.594	0	2
0: Private Independent	372	5.66%			
1: Private Government-dependent	1,859	28.27%			
2: Public	4,344	66.07%			
Clsize	6,709	27.163	7.385	13	53

Level 1 equation (student-level)

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} X_{kij} + r_{ij} \\ &= \beta_{0j} + \beta_{1j}(\text{PA003Q02TA}) + \beta_{2j}(\text{female}) + \beta_{3j}(\text{repeat}) \\ &\quad + \beta_{4j}(\text{immig}) + \beta_{5j}(\text{escs}) + \beta_{6j}(\text{ISCED0}) + r_{ij} \end{aligned}$$

$$r_{ij} \sim N\left(0, \sigma^2\right) \tag{13.1}$$

Level 2 equation (school-level)

$$\beta_{kj} = \gamma_{k0} + \sum_{q=1}^{Q} \gamma_{kq} W_{qj} + u_{kj} = \gamma_{k0} + \gamma_{k1}(\text{schltype}) + \gamma_{k2}(\text{clsize}) + u_{kj}$$

$$u_{kj} \sim N(0, \tau_1) \tag{13.2}$$

where Y_{ij} represents the score achieved by student "i" at school "j". X is a set of "k" characteristics of student "i" in school "j" (variables of level 1). β_{0j} and β_{kj} are level 1 estimated coefficients and r_{ij} are the level 1 random effects. B_{1j} is the coefficient of PA003Q02TA measuring the relationship between the frequency of shared meals and student at school "j" academic achievement. Each of the level 1 coefficients turns into a dependent variable in the level 2 equation. W_{qj} is a vector of "q" characteristics of school "j". γ_{k0} and γ_{kq} are level 2 coefficients and u_{kj} are the random effects at level 2. γ_{10} is the average effect of the frequency of shared meals in the school.

Equation (13.3) is obtained by substituting in Eq. 13.1 (student level) the coefficients of Eq. 13.2 (school level):

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} W_{qj} + \sum_{k=1}^{n} \beta_{kj} X_{kij} + r_{ij} + u_{0j} \\ &= \gamma_{00} + \gamma_{01}(\text{schltype}) + \gamma_{02}(\text{clsize}) + \beta_{1j}(\text{PA003Q02TA}) \\ &\quad + \beta_{2j}(\text{female}) + \beta_{3j}(\text{repeat}) + \beta_{4j}(\text{immig}) \\ &\quad + \beta_{5j}(\text{escs}) + \beta_{6j}(\text{ISCED0}) + r_{ij} + u_{kj} \end{aligned} \tag{13.3}$$

The models have been estimated using the mixed function of the statistical software *Stata 14*. The *mixed* command allows the application of the sampling weights for students and schools provided by the 2015 PISA database. The application of these weights allows to correct for imperfections in the sample that may lead to biases and significant differences between the sample and the reference population.

13.4 Results

The results obtained by applying multilevel techniques (Table 13.3) suggest that a higher frequency of shared family meals is associated with higher levels of performance in reading literacy. This result is in line with Cullen and Baranowski (2000), Eisenberg et al. (2004), Neumark-Sztainer et al. (2010), and Shin et al. (2017). Specifically, we find that students who share daily or almost daily their mean meal with their parents see their reading comprehension score increase by 35.46 points - when controlling all other variables as constant- compared to students who share meals at most twice a month with their parents. However, in the rest of the competences the coefficients obtained are not statistically significant. The discussion section reflects on this differential outcome by competences. The estimated variance of the constant per student (the individual level random effect) is non-zero in all the competences and that is evidence that the random effect is beneficial and multilevel modeling is appropriate.

In addition to the main estimates, we use the *margins* and *marginsplot* commands to estimate the partial effects of the variable of interest in this research: PA003Q02TA. The objective is to determine whether the positive relationship between the frequency of shared meals with the families and academic performance in reading differs according to the gender and socio-economic and cultural status (ESCS) of the student.

The results in Table 13.4 and Graph 13.1 show that for male respondents, the difference in predictive margins for PA003Q02TA are similar (493.5274 − 458.0703 = 35.4571) than for female respondents (516.4829 − 481.0257 = 35.4572). This shows that there are no differential effects by gender and that family meals improve reading achievement in similar proportions for boys and girls. In Table 13.5 and Graph 13.2 we get similar results for the index of economic, social, and cultural status (ESCS). Specifically, the analysis of the marginal effects at the nine percentile values of the ESCS variable shows that the magnitude of the positive association between the frequency of family meals and academic performance in reading comprehension is maintained regardless of the student's ESCS value.

13.5 Conclusions

The results obtained in this research show that a higher frequency of family meals is associated with better levels of academic performance in reading, although it does not seem to have any effect on performance in science, mathematics, financial literacy, and collaborative problem solving. This differential impact by competences makes sense given that, as Fruh et al. (2011) show, eating meals with families has very beneficial effects on children's and adolescents' vocabulary. A larger vocabulary has been shown to help reading skills (Snow & Beals, 2006) and this would explain why these students end up achieving higher levels of academic performance in reading comprehension, as evidenced by our results. This result, which indicates a positive

Table 13.3 Results of multilevel regressions

Variables	(1)	(2)	(3)	(4)	(5)
	Mathematics	Reading	Science	Financial	Collaborative
PA003Q02TA = 1	−10.96	28.91*	1.124	7.677	−8.606
(Once or twice a week)	(14.28)	(17.29)	(15.41)	(18.75)	(18.49)
PA003Q02TA = 2	−0.801	35.46**	3.257	−0.207	−2.098
(Every day or almost)	(12.78)	(15.90)	(14.04)	(17.48)	(16.84)
Female	−23.14***	16.59***	−14.72***	0.336	17.33***
	(2.807)	(3.052)	(2.713)	(3.386)	(3.360)
Repeat	−82.47***	−75.54***	−82.49***	−79.00***	−71.85***
	(3.516)	(3.468)	(3.638)	(4.648)	(3.800)
Immig = 1	−16.22*	−0.136	−20.83	−16.18	−19.91
(Second-Generation)	(9.574)	(12.95)	(14.31)	(17.57)	(13.28)
Immig = 2	−1.511	−0.334	−6.248	3.917	7.479
(First-Generation)	(6.576)	(5.002)	(5.892)	(7.064)	(6.556)
ESCS	9.806***	10.92***	10.89***	9.282***	5.275***
	(1.424)	(1.365)	(1.261)	(1.558)	(1.717)
ISCED0 = 1	14.12**	15.82***	14.35**	15.01*	13.51**
(2 years)	(5.588)	(4.812)	(5.848)	(7.789)	(6.319)
ISCED0 = 2	4.293	10.82**	8.797	13.67*	10.53*
(3 years)	(6.478)	(5.091)	(6.601)	(7.045)	(5.992)
ISCED0 = 3	−3.081	6.297	−4.627	−13.26	−7.278
(4 years)	(11.95)	(12.52)	(12.94)	(12.22)	(16.40)
ISCED0 = 4	−21.35*	−7.050	−18.01	−20.06	−26.94**
(5 years)	(12.98)	(10.89)	(13.93)	(15.94)	(11.00)
ISCED0 = 5	−41.17***	−40.51**	−31.87***	−21.13	−41.59***
(6 years)	(11.84)	(18.17)	(10.52)	(18.75)	(9.902)
ISCED0 = 6	−11.98	6.947	−2.904	4.634	−2.559
(did not attend)	(14.33)	(16.07)	(16.75)	(15.31)	(19.80)
Schltype = 1	−8.381	7.583	4.323	5.045	12.01
(Private Government-dependent)	(9.006)	(11.62)	(10.10)	(14.81)	(11.43)
Schltype = 2	−1.112	2.425	8.046	6.828	11.39
(Public)	(8.141)	(11.41)	(9.565)	(14.31)	(11.01)
Clsize	0.391	0.466	0.112	−0.221	0.553
	(0.398)	(0.427)	(0.361)	(0.521)	(0.397)
Constant	520.9***	459.1***	517.7***	495.1***	485.6***
	(17.97)	(21.85)	(18.44)	(26.49)	(22.13)

(continued)

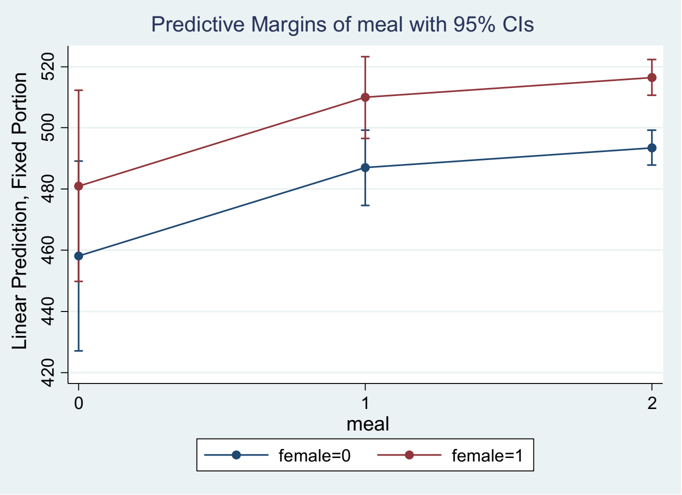
Table 13.3 (continued)

Variables	(1)	(2)	(3)	(4)	(5)
	Mathematics	Reading	Science	Financial	Collaborative
Observations	4,281	4,281	4,281	4,281	4,281
Number of groups	193	193	193	193	193
var(_cons)	769.352	866.721	779.528	1559.771	1076.917
	(127.831)	(120.693)	(106.588)	(191.160)	(120.716)

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 13.4 Predictive margins for *PA003Q02TA* by gender of the student

Female	PA003Q02TA	Margin	Std. Err.	*z*	*P* > *z*	[95% Conf. Interval]	
0	0	458.0703	15.83497	28.93	0	427.0343	489.1063
0	1	486.9812	6.260298	77.79	0	474.7113	499.2512
0	2	493.5274	2.933919	168.21	0	487.7771	499.2778
1	0	481.0257	15.9496	30.16	0	449.7651	512.2864
1	1	509.9367	6.805354	74.93	0	496.5984	523.2749
1	2	516.4829	2.999579	172.19	0	510.6038	522.362



Graph 13.1 Predictive margins for *PA003Q02TA* by gender of the student

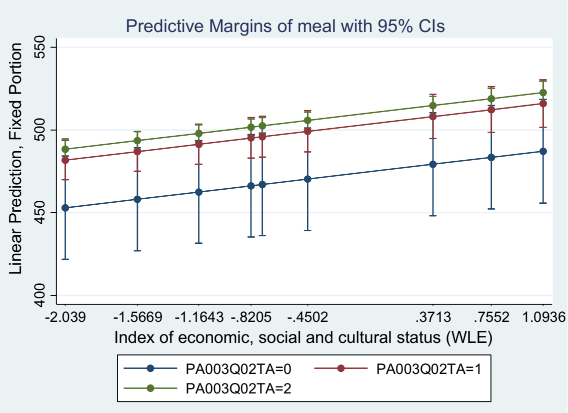
effect of a higher frequency of family meals on academic performance in reading comprehension, is in line with those obtained in previous studies (Cullen & Baranowski, 2000; Eisenberg et al., 2004; Neumark-Sztainer et al., 2010; Shin et al., 2017).

Table 13.5 Predictive margins for *PA003Q02TA* by percentile of ESCS of the student

ESCS	PA003Q02TA	Margin	Std. Err.	*Z*	*P* > *z*	[95% Conf. Interval]	
1	0	452.9221	15.9546	28.39	0	421.6517	484.1926
1	1	481.833	6.085141	79.18	0	469.9064	493.7597
1	2	488.3793	3.185206	153.33	0	482.1364	494.6222
2	0	458.0774	15.88607	28.84	0	426.9413	489.2135
2	1	486.9884	6.095225	79.9	0	475.0419	498.9348
2	2	493.5346	2.848303	173.27	0	487.952	499.1171
3	0	462.4738	15.84811	29.18	0	431.4121	493.5355
3	1	491.3847	6.157336	79.8	0	479.3166	503.4529
3	2	497.9309	2.652477	187.72	0	492.7322	503.1297
4	0	466.2281	15.83073	29.45	0	435.2004	497.2557
4	1	495.139	6.248269	79.24	0	482.8926	507.3854
4	2	501.6852	2.567874	195.37	0	496.6523	506.7182
5	0	470.2717	15.82757	29.71	0	439.2503	501.2932
5	1	499.1827	6.38347	78.2	0	486.6713	511.694
5	2	505.7289	2.571199	196.69	0	500.6894	510.7683
6	0	467.0307	15.82882	29.51	0	436.0068	498.0546
6	1	495.9416	6.272096	79.07	0	483.6485	508.2347
6	2	502.4878	2.560608	196.24	0	497.4691	507.5065
7	0	479.2425	15.87811	30.18	0	448.122	510.363
7	1	508.1534	6.808964	74.63	0	494.8081	521.4987
7	2	514.6996	2.910932	176.82	0	508.9943	520.4049
8	0	483.4347	15.92877	30.35	0	452.2148	514.6545
8	1	512.3456	7.06028	72.57	0	498.5077	526.1835
8	2	518.8918	3.194723	162.42	0	512.6303	525.1533
9	0	487.13	15.98754	30.47	0	455.795	518.465
9	1	516.0409	7.305855	70.63	0	501.7217	530.3601
9	2	522.5871	3.491293	149.68	0	515.7443	529.4299

Although the results are in line with previous research in suggesting beneficial effects on academic performance of increasing the frequency of family meals, compared to previous studies, the results obtained represent an important novelty for various reasons. Firstly, the results are novel and relevant because they place the focus of attention on reading competence and show that in the rest of the competences evaluated it is not possible to speak of an association between the frequency of family meals and academic performance. This contribution is relevant, given that most previous studies took as a variable to measure academic performance the average GPA (or school grades) across all subjects of the students, which made it impossible to distinguish whether the impact of a higher frequency of family meals

Graph 13.2 Predictive margins for *PA003Q02TA* by percentile of ESCS of the student



was different according to the subject or competence assessed. Moreover, it is also novel that we show that this positive association between the frequency of family meals and academic performance is of similar magnitude regardless of the gender or socio-economic and cultural status of the student. Additionally, the application of multilevel techniques allows us to overcome many of the most important methodological limitations of previous research in this area. The models estimated in this research consider the hierarchical structure that educational data, such as the PISA data used in this research, generally present and correct for the fact that students in the same school are not independent and thus, compared to OLS models, lead to unbiased estimates of standard errors.

Regarding the analysis of the marginal effects, the fact that the positive association is similar regardless of gender and socio-economic and cultural status is a relevant result, since it shows that increasing the frequency of family meals would benefit the academic performance in reading comprehension of different types of students, so that the positive effects of family meals are not limited to a specific group of students. This highlights the relevance of developing policies and measures aimed at increasing the frequency of family meals at home since boys and girls and students from different socio-economic and cultural backgrounds would benefit from it.

Obviously, as with all non-experimental research, our study is not without limitations, and further experimental research would be needed to confirm that the relationship we have found is causal. In addition, we believe that future research could allow us to further investigate the effects of family meals by analyzing not only the impact of their frequency, but also the type of family meals using data that provide more detailed information about this time of day, something that unfortunately the PISA 2015 data do not allow. In this sense, it may be that future research could identify habits or behaviors in family meals that are particularly favorable to children and adolescents.

Without losing sight of the limitations of the research, by way of conclusion it can be concluded that the research carried out shows that there is a positive association between a higher frequency of family meals and the academic performance achieved in reading comprehension competence in the PISA 2015 tests by Spanish students. In addition to this beneficial effect on academic performance, there are also the positive effects of family meals explained in the introduction on aspects such as family closeness and the emotional well-being of all family members (Satter, 1986), family connectedness (Fiese et al., 2002), parental well-being (Jacob et al., 2008), nutrition habits (Anderson & Whitaker, 2010; Burgess-Champoux et al., 2009; Fulkerson et al., 2009; Neumark-Sztainer et al., 2010), and prevention of high-risk behaviors (Eisenberg et al., 2004; Fulkerson et al., 2006; Neumark-Sztainer et al., 2010; Sen, 2010).

It seems clear, therefore, that family meals bring many benefits to children and adolescents and should be encouraged and facilitated. In this sense, we believe that there are two main actions that could be carried out to ensure and facilitate that students can enjoy this time with their families: (1) social awareness-raising measures about the benefits of family meals, for example through television advertisements or social media. In this way, those who choose not to eat family meals even though they are able to do so could change their perception and modify their behavior in this respect. On the other hand, (2) we believe that the main reason for not eating meals as a family is related to parents' work and incompatibility of schedules. In this sense, we consider that it is necessary to continue improving measures aimed at reconciling work and family life in order to coordinate school and work schedules. In summary, it is evident that family meals are an important moment that implies enormous benefits in the personal and academic development of children and adolescents and therefore it is necessary to encourage and facilitate a greater frequency of them.

References

Alker, H. (1969). A typology of ecological fallacies. *Quantitative Ecological Analysis in the Social Sciences, 1969*, 69–86.

Anderson, S. E., & Whitaker, R. C. (2010). Household routines and obesity in US preschool-aged children. *Pediatrics, 125*(3), 420–428.

Burgess-Champoux, T. L., Larson, N., Neumark-Sztainer, D., Hannan, P. J., & Story, M. (2009). Are family meal patterns associated with overall diet quality during the transition from early to middle adolescence? *Journal of Nutrition Education and Behavior, 41*(2), 79–86.

Choi, A., & Calero, J. (2012). Academic performance and school ownership in Spain. *Revista de Currículum y Formación del Profesorado, 16*(3), 31–57.

Cullen, K. W., & Baranowski, T. (2000). Influence of family dinner on food intake of 4th to 6th grade students. *Journal of the American Dietetic Association, 100*, A38.

Eisenberg, M. E., Olson, R. E., Neumark-Sztainer, D., Story, M., & Bearinger, L. H. (2004). Correlations between family meals and psychosocial well-being among adolescents. *Archives of Pediatrics & Adolescent Medicine, 158*(8), 792–796.

Fiese, B. H., Tomcho, T. J., Douglas, M., Josephs, K., Poltrock, S., & Baker, T. (2002). A review of 50 years of research on naturally occurring family routines and rituals: Cause for celebration? *Journal of Family Psychology, 16*(4), 381.

Fruh, S. M., Fulkerson, J. A., Mulekar, M. S., Kendrick, L. A. J., & Clanton, C. (2011). The surprising benefits of the family meal. *The Journal for Nurse Practitioners, 7*(1), 18–22.

Fulkerson, J. A., Kubik, M. Y., Story, M., Lytle, L., & Arcan, C. (2009). Are there nutritional and other benefits associated with family meals among at-risk youth? *Journal of Adolescent Health, 45*(4), 389–395.

Fulkerson, J. A., Story, M., Mellin, A., Leffert, N., Neumark-Sztainer, D., & French, S. A. (2006). Family dinner meal frequency and adolescent development: Relationships with developmental assets and high-risk behaviors. *Journal of Adolescent Health, 39*(3), 337–345.

Gómez-Fernández, N., & Mediavilla, M. (2021). Exploring the relationship between Information and Communication Technologies (ICT) and academic performance: A multilevel analysis for Spain. Socio-Economic Planning Sciences, 77. https://doi.org/10.1016/j.seps.2021.101009

Hox, J. (1995). *Applied multilevel analysis*. TT-publikaties.

Ikeda, M., & García, E. (2014). Grade repetition: A comparative study of academic and non-academic consequences. *OECD Journal: Economic Studies, 2013*(1), 269–315.

Jacob, J. I., Allen, S., Hill, E. J., Mead, N. L., & Ferris, M. (2008). Work interference with dinnertime as a mediator and moderator between work hours and work and family outcomes. *Family and Consumer Sciences Research Journal, 36*(4), 310–327.

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review, 61*, 51–58.

Kashkary, S. Y. (2012). "Are two better than one?" The Impact of Length of Time Spent in Kindergarten on Pupils' Mathematics Achievement of Primary School: a case study of grade one pupils in jeddah, Saudi Arabia. *Journal of Arabic and Human Sciences, 5*(1).

Kim, E. J., & Lee, J. (2021). The multiple dimensions of family meals and their associations with family strengths from the perspective of Korean mothers with school-aged children. *Family and Environment Research, 59*(2), 169–183.

Koc, N., & Celik, B. (2015). The impact of number of students per teacher on student achievement. *Procedia-Social and Behavioral Sciences, 177*, 65–70.

Lee, J. Y., Lee, S., Park, E. C., Kim, J., & Jang, S. I. (2020). The association of maternal accompaniment at family dinners and depressive symptoms of Korean adolescents. *International Journal of Environmental Research and Public Health, 17*(5), 1743.

Middleton, G., Golley, R., Patterson, K., Le Moal, F., & Coveney, J. (2020). What can families gain from the family meal? A mixed-papers systematic review. *Appetite, 153*, 104725.

Miller, D. P., Waldfogel, J., & Han, W. J. (2012). Family meals and child academic and behavioral outcomes. *Child Development, 83*(6), 2104–2120.

Neumark-Sztainer, D., Larson, N. I., Fulkerson, J. A., Eisenberg, M. E., & Story, M. (2010). Family meals and adolescents: What have we learned from Project EAT (Eating Among Teens)? *Public Health Nutrition, 13*(7), 1113–1121.

OECD. (2015). *Technical report PISA 2015*. OECD Publishing.

OECD. (2009a). *PISA data analysis manual*. OECD Publishing.

OECD. (2009b). *PISA data analysis manual: SAS*. OECD Publishing.

Parker, P. D., Van Zanden, B., & Parker, R. B. (2018). Girls get smart, boys get smug: Historical changes in gender differences in math, literacy, and academic social comparison and achievement. *Learning & Instruction, 54*, 125–137. https://doi.org/10.1016/j.learninstruc.2017.09.002

Robbin, J. B. (1996). *The effectiveness of preschool education on academic achievement*. ERIC. Accession No. ED, 400069.

Rutkowski, L., Rutkowski, D., & Liaw, Y. L. (2019). The existence and impact of floor effects for low-performing PISA participants. *Assessment in Education: Principles, Policy & Practice, 26*(6), 643–664.

Säälik, Ü., Nissinen, K., & Malin, A. (2015). Learning strategies explaining differences in reading proficiency. Findings of Nordic and Baltic countries in PISA 2009. *Learning and Individual Differences, 42*, 36–43.

Satter, E. M. (1986). The feeding relationship. *Journal of the American Dietetic Association, 86*(3), 352–356.

Sen, B. (2010). The relationship between frequency of family dinner and adolescent problem behaviors after adjusting for other family characteristics. *Journal of Adolescence, 33*(1), 187–196.

Shin, W. K., Kang, S. Y., & Kim, Y. (2017). Effects of family meals on eating behavior, academic achievement and quality of life-Based on the students of middle school at Goyangsi, Gyeonggido. *Journal of Korean Home Economics Education Association, 29*(4), 149–159.

Snijders, T. (2011). Multilevel analysis. In *International Encyclopedia of Statistical Science* (pp. 879–882). Springer Berlin Heidelberg.

Snow, C. E., & Beals, D. E. (2006). Mealtime talk that supports literacy development. *New Directions for Child and Adolescent Development, 2006*(111), 51–66.

Suárez-Álvarez, J., Fernández-Alonso, R., & Muñiz, J. (2014). Self-concept, motivation, expectations, and socioeconomic level as predictors of academic performance in mathematics. *Learning and Individual Differences, 30*, 118–123.

Thorpe, G. (2006). Multilevel analysis of PISA 2000 reading results for the United Kingdom using pupil scale variables. *School Effectiveness and School Improvement, 17*(1), 33–62.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461.

Nerea Gómez-Fernández is a Ph.D. candidate and researcher at the Polytechnic University of Valencia and associate professor at EDEM Escuela de Empresarios. She has also been a Ph.D. visiting researcher at the London School of Economics and Political Sciences. She completed a Bachelor's degree in Economics from the University of Valencia (National Award for Academic Excellence) and a Master's Degree in Applied Economics from the University of Alicante. Her research interests are mainly in the areas of economics of education and applied macroeconometrics.

Juan-Francisco Albert holds a Ph.D. in Applied Economics from the University of Valencia. He has been a Ph.D. visiting researcher at the London School of Economics and Political Science (LSE) and De Nederlandsche Bank. His research interest is mainly in the areas of applied macroeconomics, economic policy, and public economics.

Chapter 14
Multilevel Modeling of Nordic Students' Mathematics Achievements in TIMSS 2019

Marie Wiberg

Abstract The overall aim was to model Nordic students' mathematics achievement and try to identify factors within the schools that contribute to the explanation why schools are either low or high effective. Effective schools are defined as schools contributing efficiently to students' mathematics achievement. Three Nordic countries, Finland, Norway and Sweden, were included as they all took part in TIMSS mathematics for grade 8 in 2019. The used school factors were constructed from the TIMSS school questionnaire, which is answered by the principals of the participating schools. Multilevel analyses were used to separate the school effects from the students' home background. Not surprisingly, as the countries have different educational systems, different factors were of different importance in the different countries and for different types of schools. The identified school factors in each country were spread among context and climate factors. The practical implications of the obtained results are discussed as well as directions for future research.

Keywords Effective schools · Mathematics achievement · School level factors · Home background factors · Multilevel analysis

14.1 Introduction

Trends in Mathematics and Science Study (TIMSS) is an international large-scale assessment (ILSA) given every fourth year which aim to measure trends in students' achievement in mathematics and science in grade 8 and grade 4 around the world. Results from TIMSS 2019 indicate that the participating Nordic countries, Finland, Norway and Sweden, had the same (Norway and Sweden) or similar (Finland) average mathematics achievement but their trends differed. Sweden had similar average mathematics achievement as in 2015, Norway exhibited a decline from 2015 and Finland, which did not participate in 2015 had a decline from 2011 (Mullis et al., 2020).

M. Wiberg (✉)
Department of Statistics, USBE, Umeå University, Umeå, Sweden
e-mail: marie.wiberg@umu.se

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_14

One way to understand differences in mathematics achievement between the countries is to examine the efficiency of the schools, that is to what extent schools are successful to provide good educational opportunities to enhance educational achievement and development. It is however not as simple that one can just compare the average mathematics achievement as there are several factors which can influence the outcome. Students from homes where one emphasizes the importance of education, and have the resources needed typically perform better and thus achieve higher results. A school which provide good resources in terms of staff, space, equipment tends to have higher results. The fact that these factors play a significant role, make it evident that there are differences within and between schools.

School effectiveness research is especially focused on differences between schools, as compared with other educational effectiveness research which may also include economical or instructional studies (Scheerens & Bosker, 1997). Different aspects of school effectiveness in order to improve students' educational outcome has been studied over a long period of time according to Sammons (2007). Although definitions of school effectiveness may vary, a school is typically defined as effective if it adds extra value to the students' achievements. This means that an effective school can improve the students' achievement regardless of the characteristics of the student body. Rumberger and Palardy (2004) noted that an improvement of school effectiveness research was the start of using multilevel modeling to estimate the effects of factors on students' achievement more accurately as one could examine the effect on different levels in the education system (e.g., student level, class level and school level). A particular focus is on school aspects which can be influenced to improve students' achievements. This means that research on school effectiveness focuses on the school as the major unit of change in educational reform (Teddlie, 2010).

In this study, a school is defined as effective if it adds value to the students in terms of outcomes, regardless of the characteristics of the students within the school. In other words, if all schools had the same kind of students with respect to their initial knowledge level and preparation, then school effectiveness can be measured by comparing the student achievements at the end of the school year. This means that in order to examine school effectiveness, it is important to take care of the students' characteristics which are correlated with students' success (Martin et al., 2000).

An important aspect of school effectiveness research is to examine environmental factors, which can be categorized as climate and context variables (Ma et al., 2008). Climate variables are the school culture or "software" of a school, and refer to the administrative policies, the values and expectations of students, parents and educational professionals. The context variables are the setting or the "hardware" of the school, such as school location, resources, and the characteristics of the teacher and student body. The principal has a key role as s/he is responsible for the implementations in a school.

In general, few research studies have focused on school principals in relation to student achievements, using large-scale quantitative data (Johansson & Bredeson, 2011). There has however been some previous research on school effectiveness in the Nordic countries. Wiberg et al. (2013) examined school factors in relation to students' mathematics achievement on three TIMSS assessments (2003, 2007, and 2011) in Norway and Sweden. Wiberg and Rolfsman (2013) examined students' science achievement and the influence of different school factors using TIMSS 2003 and TIMSS 2011 in Norway and Sweden. A problem when analyzing and comparing different TIMSS assessments is that the questionnaires change substantially over the years and thus it is challenging to construct school factors which are valid over several administrations and between different countries as national adaptions also occur. In this study we only use TIMSS variables, but if one has access to more information from the school system it has been shown to add important information about the association of the students' background and TIMSS achievements both in mathematics and science (Wiberg, 2019; Wiberg & Rolfsman, 2019).

The overall aim with this study was to model Nordic students' mathematics achievement and try to identify factors within the schools that contribute to the explanation why schools are either low or high effective. As the countries have slightly different educational contexts it is likely that the school factors differ. A particular interest is to examine if there are any school factors that can be identified which are possible to influence such as school climate factors.

This study creates the opportunity to separate the school effect from the effect of students' home environment, which is important because of the well-known association between success at school and social background factors (e.g. Giddens, 1997), as background factors can affect young children's cognitive skills. The current study contributes to an examination of school factors associated with student success in the Nordic countries. As the focus was on one TIMSS cycle the reliability is strengthened as it allows the usage of all available variables. The specific research question included were:

1. Which school level factors are associated with mathematics achievement in the participating Nordic countries in TIMSS 2019?
2. Are the identified school level factors the same or do they differ for low or high effective schools?
3. Are the identified school level factors specific to a country or similar between countries?

14.2 Method

14.2.1 Participants

Data from TIMSS 2019 mathematics for 8th grade students and their schools (IEA, 2021) was used from Finland, Norway and Sweden as these were the only Nordic countries that participated in 2019 with 8th grade students. We used data from students' mathematics achievement, the students' questionnaire and the school questionnaires. The school questionnaires were answered by the school principals and were used as measures of the context and climate of the school. The school questionnaire includes questions about school enrolment and characteristics, instructional time, resources and technology, school emphasis on academic success, school discipline and safety, and principal experience and education. The students' questionnaires include questions about the student, the students' parents' educational background, the students' home possessions, the students' behavior at school, what they think about mathematics and science and questions about homework. The student questionnaire was used to control for the students' home background.

14.2.2 Statistical analyses

Only a limited number of the TIMSS mathematics achievement items are administered to each student, to limit their time and effort. The students' scores on the obtained TIMSS items are transformed into five plausible values, which represent the students' mathematics achievements on the whole assessment if they would have answered all the mathematics items. In the statistical analyses, we used the five plausible values as representing mathematics achievement and followed the suggested guidelines for TIMSS 2019 (Martin et al., 2019) and other researchers' suggestions on how to use plausible values in secondary analyses (Laukaityte & Wiberg, 2017; von Davier et al., 2009). Initially, the data files were prepared and analyzed with IEA IDB analyzer 4.0 and SPSS 26.0. The statistical analyses were then conducted in four steps;

1. Identifying student home background variables.
2. Deciding which schools were low and high effective.
3. Constructing school factors.
4. Conducting multilevel analysis.

In the first step, student home background variables were chosen from the students' questionnaires based on previous studies (Wiberg & Rolfsman, 2021) and availability of these variables in TIMSS 2019 in the three countries. The student variables' potential influence on mathematics achievement were examined within each country with multiple regressions with the five plausible values representing mathematics achievement as dependent variables. Number of home study support was not explicitly used because most students (85–97%) had access to most of them including internet, study desk, computer and a mobile phone. The number of books at students' home and the TIMSS defined home educational resource [HER] index was examined, as these two has been shown to well represent students' socioeconomic status (Wiberg & Rolfsman, 2021) and have relatively low amount of missing values. HER consists of number of books at home, highest level of education of either parent or guardian and number of home study supports (internet connection + own room). As these two variables were highly correlated we chose to use HER as it had slightly fewer missing values in all three countries. We also constructed a migration indicator variable [GB] with value 1 if at least one guardian was born within the examined country, and 0 otherwise. Missing data was in general low in the student's home background, ranging from 1% for HER (Finland) to 9% for GB (Norway, which also had all the TIMSS mathematics results for these students missing) thus listwise deletion was used to exclude missing data. This choice was made as it is robust to violations of assumptions that data are missing at random or missing completely at random, and thus results in unbiased estimates of regression coefficients (Allison, 2009). Although it is theoretically better to impute missing data we made this choice because reasonably few cases were deleted. The students' home background variables were recoded to make more sense in the analyses. Books in the home were originally coded as 0–10, 11–25, 26–100, 101–200, > 200, and this was recoded as $Book = 1$ if the students' home had more than 100 books, 0 otherwise as we wanted an indicator of the homes which had a reasonable number of books. We also included the students' sex in the analyses and this variable was coded as 1 for females and 0 for males.

In the second step, effective schools were identified by using the students' average mathematics achievement and linear regression models with the described student home background variables as covariates. For each country, the mean differences between the five mathematics plausible values and the expected scores from the linear regressions were calculated. Schools were divided into three levels of effectiveness; low, mid and high effective. Schools were concluded to be high effective if they were in the top third in their country in mathematics achievement, thus if the school performed higher than predicted from the linear regression models and thus the school reached a better result than expected considering the home background of the enrolled students. A school was concluded as mid effective if they were in the middle third in their country and thus reached the expected result considering the home background of the enrolled students. A school was concluded to be low effective if they were in the bottom third within their country and thus reached a worse result than expected considering the home background of the enrolled students. In the later multilevel analyses, the low effective and high effective schools were examined separately. The mid effective schools were not examined specifically as we instead included analyses when all schools in a country were included.

In the third step, school factors were constructed from the school questionnaires and the aim was to identify school variables which could have had an impact on school effectiveness. From available school level variables in TIMSS and previous TIMSS studies we examined several potential school factors. We examined the correlation between mean mathematics difference, obtained in the second step, and the school variables to find variables which might affect the students' mathematics achievement. From these analyses, five factors were retained which included both context and climate school variables. Below, we describe the used factors briefly and how they are coded. The school factors DIS, SUC and MSR have factors in the international data base that we recoded. The parental involvement factor was constructed from three variables as they were highly correlated and Ma et al. (2008, p. 95) proposed to merge similar variables if it can be motivated theoretically. An exact definition of each factor can be obtained upon request from the author.

[DIS] School discipline problems coded as 1 (high) or 0 (low).

[SUC] School emphasis on academic success coded as 1 (high) or 0 (medium or lower).

[MSR] Instruction affected by mathematical resources coded as 1 (Affected, a lot), 0 (not affected).

[SLO] School location (Urban/Rural) coded as 1 (more than 30,000 inhabitants), 0 otherwise.

[PI] Parental involvement consists of the variables; parental commitment, parental expectation and parental support and these were added together and a high value was coded as 1 and a low value was coded as 0.

Note, using our previous categorization, School climate factors are PI, SUC and DIS, while school context factors were SLO and MSR. In order to control for the overall context of the student factors in a school we also included the aggregated mean of the student home factors,

i.e., the aggregated GB [aGB] and the aggregated HER [aHER] in the later multilevel analyses.

In the fourth step, we carried out multilevel analysis (Gelman & Hill, 2007; Snijders & Bosker, 1999) on the high effective schools, low effective schools and all schools. Multilevel analyses were chosen as it can handle TIMSS two-stage sampling design and that the probability of selecting a sample unit is unequal (Kyriakides & Charalambous, 2005). Multilevel analysis has been used to analyze TIMSS by a large number of researcher (e.g. Ersan & Rodriguez, 2020; Martin et al., 2000; Mohammadpour et al., 2015; Webster & Fisher, 2000; Wiberg & Rolfsman, 2013). The student factors were weighted with student weights and the school factors were weighted with school weights which are included in the international data base in line with the suggestions by Laukaityte and Wiberg (2018). The dependent variables consisted of the five mathematics plausible values calculated for each student as a measure of their mathematics achievement (Mislevy et al., 1992).

For the low effective schools, high effective schools and all schools within each country we fitted three types of multilevel models; a null model, a home context model with the student factors and a full model with all significant factors included. In general, we defined the multilevel models used as follows.

Level 1 (within schools):

$$Y_{ij} = \beta_{0j} + \beta_1(HER) + \beta_2(GB) + r_{ij}.$$

Level 2 (between schools):

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(H_1a) + \gamma_{02}(aHER) + \gamma_{03}(aGB)\\ &\quad + \gamma_{04}(S1) + \ldots\gamma_{08}(S5) + u_{0j}\\ \beta_{0j} &= \gamma_{00} + \mu_{0j},\ \ \beta_{1j} = \gamma_{10},\ \ \beta_{2j} = \gamma_{20}\cdots\beta_{12j} = \gamma_{12j},\end{aligned}$$

where Y_{ij} is the mathematics achievement for student $i = 1, 2 \ldots n_j$ in school $j = 1, 2 \ldots J$ β_{0j} is mean mathematics achievement of school *j*, and r_{ij} is the random error of student *i* in school *j*. Further, γ_{00} is the grand mathematics mean for all schools and μ_{0j} is the random school effect, the deviation of school *j*:s mean from the grand mean. *S1*…*S5* are the used school level factors: DIS, SUC, MSR, SLO and PI. The examined school level factors were grand mean centered as suggested by Ma et al. (2008). We examined a number of different multilevel models for each of the countries and for the different school types. Non-significant effects were removed from the multilevel models. The proportion of within and between-school variances was examined in the chosen models as high proportion of between-school variance indicates the existence of school effects (Ma et al., 2008).

14.3 Results

The average mathematics TIMSS achievement for the examined students is differentiated on different types of school as can be seen in Table 14.1. The highest average mathematics achievement appeared in Finland, while Norway and Sweden had the same average mathematics achievement. Noticeable is that all countries had somewhat similar averages in terms of mathematics achievement within the different type of schools. The standard errors in all different types of school were all reasonably low.

In this study the proportion of between-school variance explained by the chosen multilevel model had a broad range (24–80%) within countries not differentiating on effective schools and even larger (24–99%) when differentiating on effective schools as can be seen in the last row in Table 14.2. Thus, we can motivate the existence of school effects. A comparison of the used multilevel models in the three countries reveal some similarities. For example, that the HER index was always significant in all types of schools. The student variable sex and the aggregated GB factor (aGB) as well as the school level variable DIS were never significant for any type of school in any country. To have at least one guardian born within the country (GB) had a positive effect for all low effective schools and for Norwegian high effective schools. The aggregated HER (aHER) was only significant for Swedish and Finnish schools overall and for high effective schools in Sweden.

There were no significant school level factors for Norway, except for low effective schools which had parental involvement as the sole significant predictor on school

Table 14.1 Mathematics achievement differentiated on different types of schools. On average, in low effective, mid effective and high effective schools with standard errors within parenthesis

Country	Finland	Norway	Sweden
Average	509 (2.6)	503 (2.4)	503 (2.5)
Low effective	480 (3.3)	481 (2.4)	473 (3.1)
Mid effective	513 (1.7)	507 (2.1)	509 (2.8)
High effective	537 (2.3)	531 (2.7)	539 (4.0)

Table 14.2 Significant coefficients in the multilevel analyses for the low and high effective schools as well as for the whole samples (All)

		Finland			Norway			Sweden	
	Low	High	All	Low	High	All	Low	High	All
Con	481.3	536.4	508.3	478.6	537.6	505.0	473.3	539.8	505.2
HER	19.0	15.7	16.5	15.1	19.3	17.2	15.6	17.8	17.1
GB	23.0	18.7	16.8	27.5		14.2	35.3		23.0
aHER			18.6					16.1	19.4
SUC		15.4	9.2						
MSR	10.7						−10.8		−11.2
PI	−10.9	15.5		−12.3					
SLO									−9.4
BSV	0.59	0.97	0.52	0.94	0.78	0.24	0.99	0.83	0.80

HER = Home Educational Resources, GB = At least one guardian/parent was born in country, aHER = Aggregated HER, SUC = School academic success, MSR = Mathematics school resources, PI = Parental involvement, and SLO= School location, BSV = Proportion between-school variance explained

level. To attend an urban school was associated with an estimated decrease of 9 score points in mathematics when examining all schools in Sweden but not in any of the other examined cases. Interestingly, parental involvement was significant in both low and high effective schools in Finland and in low effective schools in Norway but never in Sweden. The positive coefficient for high effective schools in Finland might be explained as it is helpful in a high effective school with parental involvement. The negative coefficient for low effective schools in Norway and Finland may be due to parents are involved but not really helping in school. The non-significant coefficient in the Swedish schools may be due to very few parents are involved in any schools as that is not part of the education system.

In Finland overall, and for high effective schools the schools' emphasis of academic success was important. Finally, lack of mathematics resources was significant in low effective schools in Finland and Sweden and overall in Sweden. In the Swedish schools it was a negative coefficient suggesting that when there is a lack of resources the average TIMSS achievement is lower. In low effective schools in Finland, however, the coefficient was positive. Although this appear strange it could be that these schools try harder as they know they have shortages. The amount of shortages may also differ between countries.

14.4 Discussion

The overall aim was to model Nordic students' TIMSS mathematics achievement and try to identify factors within the schools that contribute to the explanation why schools are either low or high effective. Multilevel analyses were used, as it has improved school effectiveness research by allowing to estimate the effects of factors on student outcomes more accurately as one has the possibility to examine the effects at different levels in the education system (Rumberger & Palardy, 2004). The focus in school effectiveness research is on differences between schools (Scheerens & Bosker, 1997), and it would especially be beneficial to find school-related factors associated with students' mathematics achievement that could be influenced, as the school is the most important unit of change in educational reforms (Teddlie, 2010). The overall results were that similar home background variables were significant in all countries and regardless of the effectiveness of school, but the school level factors differed if the school was viewed as high or low effective or if all schools were examined. This is probably due to the fact that we can never remove the background of a student but different schools and different cultures develop differently.

High effective schools in Finland had parental involvement and an emphasize on academic success. For high effective schools in Norway, no significant school level factors were found and for high effective schools in Sweden only the aggregated HER index was significant. This does not mean that there are no school factors that can influence high effective schools in Norway and Sweden, only that we did not include any such factor in our study.

Low effective schools were either associated with the school factors parental involvement (Finland and Norway) or mathematical school resources (Finland and Sweden). This latter result is not surprising as previous studies has shown that school resources are related to student achievement (e.g. Bonnano & Timbs, 2004; Chan, 2008; Dustmann et al., 2003) and it has also been noted within TIMSS (e.g. Mullis et al., 2005); to only find a few significant school factors in Norway and Sweden are in line with previous studies of TIMSS data in Sweden and Norway (Wiberg & Rolfsman, 2013; Wiberg et al., 2013).

Although we only found some significant school climate variables for some countries and types of schools (i.e., parental involvement and schools emphasize on academic success), it should not be concluded that school climate does not influence students' achievement. It is common that school climate variables show no or weak effects on students' educational outcomes, especially in the presence of school context variables (p. 90; Ma et al., 2008). A limitation with the conducted study is that we did not include any information on the class level, i.e., from the teacher. In the future one should examine the influence of the teachers in order to get a more comprehensive view of the climate of the school, as the teachers meet the students every day and are important carriers of the school climate.

References

Allison, P. D. (2009). Missing data. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 72–89). Sage.

Bonnano, K., & Timbs, J. (2004). Linking school libraries to student achievement. *Independent Education, 34*, 21–22.

Chan, C. (2008). The impact of school library services on student achievement and the implications for advocacy: A review of the literature. *Access, 22*, 15–20.

Dustmann, C., Rajah, N., & van Soest, A. (2003). Class size, education and wages. *Economic Journal, 113*, F99–F120.

Ersan, O., & Rodriguez, M. C. (2020). Socioeconomic status and beyond: A multilevel analysis of TIMSS mathematics achievement given student and school context in Turkey. *Large-scale Assessments in Education, 8*(15). https://doi.org/10.1186/s40536-020-00093-y

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Giddens, A. (1997). *Sociology*. Political Press.

IEA. (2021). *TIMSS 2019 international database.* Retrieved from https://www.iea.nl/data-tools/repository/timss on April 9 2021.

Johansson, O., & Bredeson, P. V. (2011). Framtida forskningsperspektiv på rektor – vilken forskning saknas [Future research perspectives on principals: Which research is lacking? in Swedish]. In O. Johansson (red.), *Rektor en forskningsöversikt 2000–2010* [Principal—A review of research 2000–2010] (chapter 4, pp. 61–74) (Vetenskapsrådets rapportserie, 4). Vetenskapsrådet.

Kyriakides, L., & Charalambous, C. (2005). Using educational effectiveness research to design international comparative studies: Turning limitations into new perspectives. *Research Paper in Education, 20*(4), 391–412.

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large–scale assessments. *Communication in Statistics—Theory and Methods, 46*(22), 11341–11357.

Laukaityte, I., & Wiberg, M. (2018). The importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communication in Statistics—Theory and Methods, 47*(20), 4991–5012.

Ma, X., Ma, L., & Bradley, K. (2008). Using multilevel modelling to investigate school effects. In A. A. O.´Connell & D. B. McCoach (Eds.), *Multilevel modelling of educational data* (Chapter 3, pp. 59–110). Information Age Publishing.

Martin, M. O., Mullis, I. V. S., Gregory, K. D., Hoyle, C., & Shen, C. (2000). *Effective schools in science and mathematics.* IEA's third international mathematics and science study, IEA.

Martin, M. O., von Davier, M., & Mullis, I. V. S. (2019). *Methods and procedures: TIMSS 2019 technical report*. Retrieved April 29 2021 from https://timssandpirls.bc.edu/timss2019/methods/

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.

Mohammadpour, E., Shekarchizadeh, A. T., & Kalantarrashidi, A. H. (2015). Multilevel modeling of science achievement in the TIMSS participating countries. *The Journal of Eductional Research, 6*, 449–464. https://doi.org/10.1080/00220671.2014.917254

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B., (2020). *TIMSS 2019 international results in mathematics and science.* Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks.* Boston College.

Rumberger, R. W., & Palardy, G. J. (2004). Multilevel models for school effectiveness research. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 235–258). Sage.

Sammons, P. (2007). *School effectiveness and equity: Making connection, a review of school effectiveness and improvement research and its implications for practitioners and policy makers.* CfBT Education trust.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness.* Pergamon.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling.* Sage.

Teddlie, C. (2010). The legacy of the school effectiveness research tradition. In A. Hargreaves, E. Liberman, M. Fullan, & D. Hopkins (Eds.), *Second international handbook of educational change* (pp. 523–554). Springer.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 9–36.

Webster, B. J., & Fisher, D. L. (2000). Accounting for variation in science and mathematics achievement: A multilevel analysis of Australian data third international mathematics and science study (TIMSS). *School Effectiveness and School Improvement, 11*(3), 339–360.

Wiberg, M. (2019). The relationship between TIMSS mathematics achievements, grades and national test scores. *Education Inquiry, 10*(4), 328–343. https://doi.org/10.1080/20004508.2019.1579626

Wiberg, M., & Rolfsman, E. (2013). School effectiveness in science in Sweden and Norway viewed from a TIMSS perspective. *Utbildning Och Demokrati, 22*(3), 69–84.

Wiberg, M., & Rolfsman, E. (2019). The association between science achievement measures in schools and TIMSS science achievement in Sweden. *International Journal of Science Education, 41*(16), 2218–2232. https://doi.org/10.1080/09500693.2019.1666217

Wiberg, M., & Rolfsman, E. (2021). Students' self-reported background SES measures in TIMSS in relation to register SES measures when analyzing students' achievements. *Scandinavian Journal of Educational Research.* https://doi.org/10.1080/00313831.2021.1983863

Wiberg, M., Rolfsman, E., & Laukaityte, I. (2013, July 28–30). *School effectiveness in mathematics in Sweden and Norway 2003, 2007 and 2011*. Paper presented at the 5th IEA International Research Conference in Singapore.

Marie Wiberg is a Professor in Statistics with a specialty in psychometrics at Umeå University, Sweden. She has published several papers using TIMSS data and multilevel analyses. Her research interest includes international large-scale assessments, equating of test scores, estimation of ability, and psychometrics in general. She has been the associate editor of the Journal of Educational Measurement and a coordinate editor for Behaviormetrika. She has a long list of publications in different journals such as Communications in Statistics, Psychometrika, Journal of Educational and Behavioral Statistics, Journal of Educational Measurement, Applied Psychological measurement, and International Journal of Testing.

Chapter 15
Teachers' Perceptions of School Ethical Culture: The Implicit Meaning of TIMSS

Orly Shapira-Lishchinsky

Abstract This chapter aims to explore whether a shared perception of 'School Ethical Culture' (SEC) emerged from teachers' questionnaires that were distributed in 45 participant countries on behalf of the Trends in International Mathematics and Science Study (TIMSS 2015). Based on *Multiple group confirmatory analysis*, the results support a universal perception of SEC among teachers, which was elicited from teachers' responses to TIMSS questionnaires. This led to the understanding that the TIMSS teachers' questionnaire has additional meaning, which goes beyond its original factors. The results also contributed to understanding the meaning of SEC among teachers, by identifying its four dimensions: 'teachers' profession,' 'care for students' learning,' 'interaction with colleagues,' and 'respect of rules.' This may be a new measure that, up until now, has never been investigated in schools. The findings may support a universal perspective, showing how common perceptions of SEC affect student achievements. However, the different impact of SEC dimensions across countries may be explained by the national context, which depends on specific policies of each country.

Keywords TIMSS · School ethical culture · Student achievements · Teachers' profession · Care for students' learning · Interaction with colleagues · Respect of rules

15.1 Introduction

It has become increasingly important to explore teachers' perceptions of 'ethical culture' in educational systems via a cross-national perspective, since ethics constitutes an inseparable component of teaching and education all over the world (Campbell, 2011). There are two main approaches in comparative studies that focus on ethics and culture: one focuses on *different ethical perceptions* that are rooted in county culture and norms (Melé & Sánchez-Runde, 2013; Rausch et al., 2014) and the other

O. Shapira-Lishchinsky (✉)
Bar-Ilan University, Ramat Gan, Israel
e-mail: Orly.Shapira@biu.ac.il

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_15

centers on *globalization through similarities* in moral attitudes and behaviors across countries (Cullen et al., 2004; Donnelly, 2013).

This chapter was conducted with these approaches in mind while also being conscious of the fact that while there are numerous studies on teachers' perceptions of the 'ethical climate' in educational systems (e.g., Sagnak, 2010; Shapira-Lishchinsky & Raftar-Ozery, 2018), teachers' perceptions of 'ethical culture' in these systems have not been investigated enough. Moreover, we found confusion in the literature regarding these concepts, and that researchers use these concepts without differentiating between them (e.g., Denison, 1996). Furthermore, we found very few studies that investigated the differences between them (e.g., Kaptein, 2011). Thus, this chapter attempts to investigate the meaning of 'ethical culture' in schools, based on teachers' perceptions. It is also unique in its cross-national approach, focusing on teachers' perceptions from Trends in International Mathematics and Science Study (TIMSS, 2015) assessments.

This chapter had three interdependent primary goals: (a) to explore whether a shared perception, 'school ethical culture' (SEC), emerged from teachers' TIMSS questionnaires. If SEC were to be found, then our derived goals were: (b) to find the meaning of SEC, based on teachers' perceptions using a cross-national assessment; and (c) to investigate the effect of SEC teachers' perceptions on students' science achievements in the countries participating in TIMSS (2015).

Our motivation to find a *shared* meaning for the concept SEC relates to international assessments in education, such as the TIMSS, which focuses on the existence of common ethical perceptions in participating countries, such as equity and quality (Mullis et al., 2016a). In this chapter, we chose to focus on countries that participated in the TIMSS, because, for the last 20 years, TIMSS reports have exposed ethical meaning by identifying gaps in resources, opportunities, inequity and equity issues (Mullis et al., 2016b).

Below, the theoretical background that supports this chapter: ethics in the context of national and universal culture; the definitions of culture and climate in the context of ethics, the confusion around these concepts, and finally, the ethical aspects of TIMSS.

15.2 Theoretical Background

15.2.1 Ethics in the Context of National and Universal Culture

Two primary approaches focus on ethics and national culture. The first approach looks at *differences* in moral perceptions and moral judgments among cultures (Melé & Sánchez-Runde, 2013). The literature that is based on this approach finds that national culture affects ethical perceptions and behaviors in organizations (Minkov & Hofstede, 2011).

House et al. (2004), in their GLOBE (Global Leadership and Organizational Behavior Effectiveness) project, define national culture as the common experience of individuals that results in shared values, beliefs, policies and interpretations of significant events that lead to distinctive ways of perceiving the world. Minkov and Hofstede (2011) fashioned this perspective into a four-dimensional model of national culture. Their model presents a cornerstone for cross-cultural research and reflects current social values and practices (Shiraev & Levy, 2015).

We found previous studies that support the first approach. Cultural differences have been found to impact individuals' ethical reasoning skills (Christians et al., 2015). Forsyth et al. (2008) conducted a meta-analysis of data from 29 different countries and found that Western countries exhibited a more pragmatist ethic, whereas Eastern and Middle Eastern countries were more subjective and context-driven when it came to determining moral rules. Ho (2010) uncovered differences in the ethical perceptions of Malay, Chinese and Indian leaders, and indicated that differences in these cultures focus on various ethical attributes of moral dilemma. Li and Persons (2011) found that cultural differences resulted in less ethical decision-making in Chinese students, as compared to American students when focusing on an experimental corporate code of ethics.

The second approach supports the position of *universalism* vis-à-vis a perception of ethics. For example, Cullen et al. (2004) used institutional anomie theory to develop hypotheses related to four national variables of culture (achievement, individualism, universalism and pecuniary materialism). The researchers found cross-national consistency of perceptions regarding ethically suspect behaviors. Additional researchers have supported the concept of universal minimal morality, and have argued that certain basic values are necessary for collective survival (Donnelly, 2013; Ivison, 2010).

Furthermore, empirical studies have demonstrated that beyond specific moral judgment, there are basic values or principles underlying these judgments, and their common principles appear in the major world religions and traditions (Terry, 2011; Tullberg, 2015). Moreover, the universal approach may support the application of a universal ethical policy in human rights, beginning with the Universal Declaration of Human Rights and following other UN human rights covenants, along with the UN Global Compact and its 10 ethical principles (Melé & Sánchez-Runde, 2013).

In the field of education, we found studies on *cultural diversity and dissimilarit*y in different countries concerning ethical issues, such as social justice (Banks, 2015), ethical dilemmas (Milner, 2010) and the importance of developing student potential (Klassen et al., 2010). Other studies explored *globalization and similarity*, such as human rights in educational systems (Stromquist & Monkman, 2014), the reduction of gaps (Zhao, 2010) and quality education (Wang et al., 2011).

By being mindful of these two approaches in educational systems—*cultural diversity between countries* and *globalization based on common values and ethics*—we chose to undertake a cross-national study in order to learn whether countries share a common perception concerning the concept of SEC. If we were to find the existence of a shared perception for SEC, we would then try to explain students' science achievements in TIMSS countries, based on teachers' perceptions regarding SEC.

15.2.2 Confusion Around the Definitions of Culture and Climate in the Context of Ethics

We are aware that previous studies have adapted different approaches concerning the use of definitions and distinctions between culture and climate. For example, Schein (2010) considered climate as an artifact of culture and defined culture as shared norms, values and assumptions. Earlier, Denison (1996) perceived that culture and climate are not fundamentally different.

However, based on additional studies, we argue that ethical culture and ethical climate are distinct from one another. For example, Kaptein (2011) distinguished between ethical culture and ethical climate, explaining that ethical culture presents the actual conditions for ethical behaviors, while ethical climate can be defined as the expectations of stakeholders about what constitutes ethical behavior in the organization. In support of our argument, Treviño et al. (1998) found that although ethical culture and climate are highly correlated, there is a difference between the two: ethical climate relates to attitudes while ethical culture relates more to influences on behaviors. Therefore, according to their approach, ethical culture explained unethical behavior better than ethical climate did.

Considering the fact that there are numerous studies dealing with school ethical climates in comparison to studies that consider ethical culture in schools, this chapter is pioneering in its approach. It looks for a meaning for the concept of ethical culture in schools from a cross-national perspective.

15.2.3 School Ethical Culture

In an effort to understand unethical behavior in the workplace, scholars initially focused on the *personal* characteristics of individual transgressors (Treviño & Youngblood, 1990). In recent years, scholarly focus has shifted to the characteristics of the *organizational* context within which unethical behavior occurs (Kish-Gephart et al., 2010; O'Boyle et al., 2011).

'Organizational culture', often delineated by shared values (Schein, 2010), is the informal control system of an organization that comprises common traditions (Ruiz-Palomino & Martínez-Cañas, 2014). As a subset of organizational culture, the 'Ethical Culture' (EC) of an organization encompasses the expectations as to how the organization may encourage its members to behave ethically (Treviño & Weaver, 2003). Thus, EC is defined as those aspects of the perceived organizational context that may promote ethical behavior and reduce unethical behavior (Ruiz-Palomino et al., 2013).

Based on Kaptein's research (2011), EC in schools can be viewed as resulting from the interplay between the formal (e.g., educational policy) and informal (e.g., colleagues' behavior, norms concerning school ethics) systems that potentially enhance ethical behavior among teachers. In essence, based on previous studies (e.g., Kish-Gephart et al., 2010), EC relates to perceptions of what the organization is about in practice, pertaining to the conditions for ethical and unethical behavior.

Kaptain (2008) refined the construct of EC with multiple normative dimensions, focusing on ethics in terms of their virtues. He distinguished between the following virtues:

(1) '*Clarity of ethical standards*' concerns the extent to which leaders and their employers are expected to adhere to ethical standards.
(2) '*Ethical role modeling of management and supervisors*' implies the extent to which leaders and supervisors set good examples, in terms of ethics.
(3) '*Feasibility*' reflects the conditions created by the organization that can enable employees to comply with normative expectations.
(4) '*Supportability*' reflects the extent to which the organization supports ethical conduct among its leaders and employees.
(5) *'Transparency' (visibility)* is the degree to which the consequences of the conduct of leaders and their employees are perceptible.
(6) '*Discussability*' refers to the opportunity to discuss ethical issues, such as ethical dilemmas or alleged unethical behaviors. Through sharing and discussing issues, people learn from each other and are more motivated to respect each other.
(7) '*Sanctionability*' is the extent of enforcement of ethical behavior; it metes out punishment for behaving unethically and gives rewards for behaving ethically.

Based on these dimensions, Kaptain (2008) conceptualized the CEV model (the Corporate Ethical Virtues model) and developed a self-report questionnaire for measuring the ethical culture of organizations, based on different codes of ethics from around the world. The CEV model was tested and validated in different countries among managers, employees and university students (e.g., Mitonga-Monga & Cilliers, 2015; Riivari & Lämsä, 2014). However, while these studies strove to understand the meaning and validate the dimensions of EC in business and public organizations, in this chapter, we examined whether the meaning of school ethical culture could be generated from the TIMSS 2015 teachers' questionnaires. We chose this methodological approach since the TIMSS questionnaire taps teachers' perceptions of items that appear to reflect ethical conditions in schools.

15.2.4 The Ethical Aspects of TIMSS

One of the goals of TIMSS is to promote educational equity that seeks to close achievement gaps and reduce test score differences between higher and lower scoring groups (Mullis et al., 2016a). Moreover, participating countries design educational policies that take into consideration equity issues via examples, such as promoting students' potential development by maximizing the performance of low-achieving students (Hanushek & Woessmann, 2015). Thus, the ethical context that appears in the TIMSS reports encourages choosing countries that participated in TIMSS for our sample and for the elicitation of the SEC concept.

15.3 Method

15.3.1 Context

TIMSS 2015 continued a 20-year international assessment of math and science, conducted by the International Association for the Evaluation of Educational Achievement (IEA), among school principals, teachers and students. The current chapter focuses on teachers' TIMSS 2015 questionnaire responses in relation to 8th grade students' science achievements in 45 countries. To accomplish this purpose, data were based on: (a) questionnaires completed by teachers, focusing mainly on their challenges, satisfaction, professional development, and experiences in teaching, and (b) their students' achievements, based on questionnaires focusing on the science curriculum. Our analyses were based on a dataset available to all on the TIMSS website that already codes all the relevant items in the teachers' and students' questionnaires.

15.3.2 Sample

The sample was comprised of 8353 science teachers (67.7% were women) and 280,130 students (the gender proportion was equal) nested in 8353 different schools (mainly from a single class per school) across 45 countries that participated in the TIMSS 2015 survey. The majority of teachers had a Bachelor's degree or equivalent (58.2%), and the others had graduate degrees (most of them had completed a Master's level, 1.6% had completed doctoral studies). The teaching experience varied from 1–48 years with an average of about 15 years ($SD = 9.0$). The age categories indicate that 15% of the teachers were between 20 and 30, 35% were 31 and 40, 32.5% were between 41 and below 50, and the rest were up to 60 years of age. All students came from the 8th grade. We focused on the students' science scores, using the plausible value procedure (Foy, 2017). The majority of students were native born (88.9%). However, this percentage varied across countries.

15.3.3 Overview of Procedures and Analyses

Research ethics committee approval was obtained from the authors' university. In our analyses, we used SPSS V.24.0 (SPSS IBM and Corporation Released, 2017) and Mplus V.8.0 (Muthén & Muthén, 2017). The main procedures were:

Missing values. We found that out of the 8,353 teachers, 453 did not answer any of the 38 survey items that reflected SEC. In addition, 10 teachers provided partial answers (to 20 items or less). In sum, there were 463 teachers with no answers to limited answers. As a result, these teachers were excluded from the data. We then tested the missing value patterns of the remaining 7890 schools. Thus, the sample includes 94.4% of the original sample of the schools. The preliminary analysis for missing values showed that 1% or less was missing. Although the missing pattern did not exhibit clear randomness (Little & Rubin, 2014), we imputed the missing values, since the number of observations is high. For the imputation procedure, we used the expectation maximization (EM) method that improves likelihood in comparison to the known likelihood of the data (Do & Batzoglou, 2008).

Weighting. We found that the distribution of the number of schools in each country ranged from 48 schools (Malta) to 477 schools (United Arab Emirates). The analyses required that the number of schools within a country would be similar across all 45 countries. Therefore, we constructed a country weight (COUWGT) that equalized the number of schools across countries (Foy, 2017). The overall frequencies by country (Table 15.1), are presented in comparison to the weighted number of schools, where weight was calculated around the mean number of schools per country ($M = 186$). That is, when the number is lower, the weight inflates it to the mean, and if it is higher, the weight deflates it to the mean.

15.4 Results

This chapter focuses on the following steps:

The first step: Exploring whether a shared perception of 'School Ethical Culture' emerged from the teachers' questionnaires

Expert judgment. We sent emails to 10 experts in school ethical research (university professors from the TIMSS participating countries) and asked them to independently rank relevant items (questionnaire items) according to their potential SEC meaning (86 items). Each item was ranked by these experts on a scale of 1 (low) to 5 (high) in relation to their SEC relevancy. This ranking was then used to explore items which are highly relevant for SEC assessment. We followed a ranking procedure (Meyer & Booker, 2001), which recognizes high relevancy (4–5) for SEC (Appendix). Our final set of SEC items included 38 out of 86 items in the TIMSS teachers' questionnaire.

Exploratory factor analysis. Exploratory factor analysis was run on a training set (approximately one-third of the total teachers' data, $n = 2629$ teachers). Table 15.2 provides the final factor loadings for four representing factors. Out of the primary 38 items, seven were excluded due to poor loadings, loadings $< \approx 0.35$ (Osborne, 2015). The final factors were determined according to their theoretical contribution to the definition of SEC in schools, which elicited four main dimensions: 'teachers' profession', 'care for students' learning', 'interaction with colleagues', and 'respect for rules'. At the bottom of Table 15.2, we note the internal consistency (Cronbach's Alpha), which shows high internal consistency among the factor items (alpha > 0.80). The shaded cells represent the final set of items for each factor. At that exploratory point, the multilevel structure of the data is ignored.

Table 15.1 Unweighted and weighted school frequency

Country code	Country name	Unweighted school frequency	Weighted school frequency
36	Australia	285	186
48	Bahrain	105	186
72	Botswana	159	186
124	Canada	276	186
152	Chile	171	186
158	Chinese Taipei	190	186
268	Georgia	153	186
344	Hong Kong, SAR	133	186
348	Hungary	144	186
364	Iran, Islamic Republic of	250	186
372	Ireland	149	186
376	Israel	198	186
380	Italy	161	186
392	Japan	147	186
398	Kazakhstan	172	186
400	Jordan	252	186
410	Korea, Republic of	150	186
414	Kuwait	168	186
422	Lebanon	138	186
440	Lithuania	208	186
458	Malaysia	207	186
470	Malta	48	186
504	Morocco	345	186
512	Oman	301	186
554	New Zealand	145	186
578	Norway	143	186
634	Qatar	131	186
643	Russian Federation	204	186
682	Saudi Arabia	143	186
702	Singapore	167	186
705	Slovenia	148	186
710	South Africa	292	186
752	Sweden	150	186
764	Thailand	204	186
784	United Arab Emirates	477	186

(continued)

Table 15.1 (continued)

Country code	Country name	Unweighted school frequency	Weighted school frequency
792	Turkey	218	186
818	Egypt	211	186
840	United States	246	186
926	England	143	186
5788	Norway -8	142	186
7841	United Arab Emirates (Dubai)	135	186
7842	United Arab Emirates (Abu Dhabi)	156	186
9132	Canada (Ontario)	138	186
9133	Canada (Quebec)	122	186
32,001	Argentina, Buenos Aires	128	186
Total			8353

Multilevel confirmatory analysis. Our exploratory analyses (above) led to a four-dimensional factor structure, which represents the multidimensionality of teachers' perceptions of the SEC. In the confirmatory modeling approach, we aimed to confirm this factor structure. A question arose as to the multilevel arrangement of these factors. Do factors remain the same, that is, show similar loadings for the school level and the country level? To test this possibility, we ran a *multilevel confirmatory analysis* first, and then compared the fit quality to the fit quality of a constrained model, in which loadings are held equal across the two levels. The confirmatory runs were done on the complementary set ($n = 5261$) of the data ($n = 7890$). When goodness-of-fit remains similar, that is, for example, $\Delta\text{CFI} < 0.01$, the equal loading constraint does not cause a severe reduction in the model goodness-of-fit. Therefore, it can be concluded that the factor structure at the schools' level remains similar at the country level (Heck & Thomas, 2015). Moreover, this implies that countries are similar in the overall mean teachers' SEC.

Table 15.3 presents the result of this methodology. For each original item, the intra-class correlation (ICC) coefficient was added, to test the variability which stems from the country level. We found that the ICC values were greater than 0.05 across all items; that is, a meaningful variation existed across countries as well as across schools.

The factor loadings were all high for both the within and the between levels. We tested whether factor loadings were similar across the two levels by means of measurement invariance; that is, we undertook a comparison between the configural (unconstrained) model fit and the equal loading constrained model.

The reduction in CFI between the unconstrained and the constrained model was $0.956 - 0.952 = 0.004$, for the 'teachers' profession' factor, which is lower than 0.01. Therefore, it was concluded that there was structural similarity for the

Table 15.2 Exploratory factor analysis and factor loadings ($N = 2629$ teachers[1])

	F1	F2	F3	F4
Teacher's profession				
BTBS17D Adapting my teaching to engage students' interest	0.70	0.12	−0.06	−0.08
BTBS17E Helping students appreciate the value of learning science	0.70	0.10	−0.02	−0.08
BTBS17A Inspiring students to learn science	0.65	0.10	−0.08	−0.03
BTBS17F Assessing student comprehension of science	0.64	0.08	−0.06	0.04
BTBS17C Providing challenging tasks for the highest achieving students	0.64	0.03	0.03	−0.05
BTBS17G Improving the understanding of struggling students	0.62	0.09	−0.03	0.01
BTBS17B Explaining science concepts or principles by doing science experiments	0.55	0.12	−0.05	0.00
BTBG14G Encourage students to express their ideas in class	0.54	−0.10	0.04	0.03
BTBG14D Encourage classroom discussions among students	0.53	−0.12	0.16	0.04
BTBG14C Ask students to complete challenging exercises that require them to go beyond the instruction	0.51	−0.16	0.11	0.07
BTBG14F Ask students to decide their own problem solving procedures	0.49	−0.11	0.15	0.01
BTBG14B Ask students to explain their answers	0.47	−0.18	0.03	0.08
Care for students' learning				
BTBG06P Amount of instructional support provided to teachers by school leadership	−0.10	0.86	0.06	−0.04
BTBG06O Collaboration between school leadership and teachers to plan instruction	−0.05	0.84	0.08	−0.09
BTBG06Q School leadership's support for teachers' professional development	−0.11	0.77	0.08	−0.02
BTBG06N Clarity of the school's educational objectives	0.06	0.70	−0.08	0.07
BTBG06A Teachers' understanding of the school's curricular goals	0.13	0.52	−0.06	0.04
BTBG06D Teachers working together to improve student achievement	0.05	0.47	0.19	0.02
BTBG06C Teachers' expectations for student achievement	0.15	0.36	−0.14	0.18
Interaction with colleagues				
BTBG09E Work together to try out new ideas	0.05	0.01	0.82	−0.04
BTBG09C Share what I have learned about my teaching experiences	0.01	−0.03	0.75	0.00
BTBG09F Work as a group on implementing the curriculum	0.05	0.05	0.73	0.03
BTBG09A Discuss how to teach a particular topic	0.05	−0.05	0.72	0.04
BTBG09G Work with teachers from other grades to ensure continuity in learning	−0.01	0.07	0.71	0.05

(continued)

[1] One teacher per school

Table 15.2 (continued)

	F1	F2	F3	F4
BTBG09D Visit another classroom to learn more about teaching	0.00	0.09	0.63	−0.09
Respect of rules				
BTBG07D The students behave in an orderly manner	0.01	−0.09	−0.01	0.90
BTBG07E The students are respectful of the teachers	0.04	−0.06	0.00	0.88
BTBG07F The students respect school property	0.00	−0.02	0.01	0.81
BTBG07G This school has clear rules about student conduct	−0.06	0.31	0.00	0.47
BTBG07H This school's rules are enforced in a fair and consistent manner	−0.10	0.36	0.07	0.46
BTBG06M Students' respect for classmates who excel in school	0.05	0.27	−0.03	0.39
Mean score				
STD				
Reliability—Alpha Cronbach	0.87	0.86	0.89	0.87

Item BTBG06C is excluded from the final analysis due to low loading (L < 0.35)

teachers' profession factor. In other words, we could conclude that a similar teachers' profession factor structure exists, both at the school level and the country level.

Factor 2, 'caring for learning', yielded a similar result. That is, the CFI difference equals 0.001. This was also true for the other two factors: the CFI difference for 'interaction with colleagues' was 0.002, and 0.004 for the 'respecting rules' factor. In sum, the conclusion is that the factor structure, as observed within each country (school level), was also found between countries (country level).

Multiple group confirmatory analysis. Another approach to confirm similar dimensionality of factors across countries is the implementation of a multiple group analysis by a measurement invariance test. In this method, we first allowed free loadings on factors for each country. That is, we used an independent measurement model (configural model) across the different countries. This step provides a reference for the overall goodness-of-fit. The measurement invariance test is a gradual imposition of loading constraints across countries to assess fit reduction, due to these constraints (Scmitt & Kuljanin, 2008). We gradually imposed equal loadings and then equal loadings and intercepts across all countries.

The first constraint, metric stage, tested whether the factor structure (regression slopes) was equal across all countries. The stricter constraint model (Scalar model) tested whether the structure and the level (intercepts) of the factors differed across countries. Similar to the multilevel test, if fit indices demonstrated minor reduction, the conclusion would be that all countries shared a similar factor structure. However, if the reduction in goodness-of-fit was significant (ΔCFI > 0.01), it would indicate a different structure across countries and would require a finer comparison, to detect the countries which were different.

Table 15.3 The multilevel confirmatory model results—Factor loadings and invariance test (N = 5261 teachers[2])

Factor					
Teacher's profession	Within level		Between level		ICC
	Loadings	SE	Loadings	SE	
BTBS17D	0.74***	0.01	0.97***	0.02	0.11
BTBS17E	0.73***	0.01	0.93***	0.04	0.14
BTBS17A	0.67***	0.02	0.83***	0.07	0.18
BTBS17F	0.66***	0.01	0.84***	0.07	0.09
BTBS17C	0.66***	0.02	0.83***	0.05	0.10
BTBS17G	0.63***	0.01	0.76***	0.09	0.10
BTBS17B	0.63***	0.01	0.59***	0.11	0.12
BTBG14G	0.40***	0.02	0.73***	0.07	0.15
BTBG14D	0.37***	0.02	0.72***	0.09	0.21
BTBG14C	0.36***	0.02	0.64***	0.10	0.17
BTBG14F	0.40***	0.02	0.63***	0.08	0.18
BTBG14B	0.30***	0.02	0.63***	0.12	0.13
Unconstrained Model fit: CFI = 0.956, TLI = 0.940, RMSEA = 0.035, Chi Square = 722.43, df = 97, $p < 0.001$, AIC = 100,691.91, BIC = 101,158.24 Constrained Model fit: CFI = 0.952, TLI = 0.942, RMSEA = 0.035, Chi Square = 792.85, df = 109, $p < 0.001$, AIC = 100,730.67, BIC = 101,118.19					
Care for students' learning	Within level		Between level		ICC
	Loadings	SE	Loadings	SE	
BTBG06P	0.83***	0.01	0.97***	0.02	0.14
BTBG06O	0.86***	0.01	0.94***	0.03	0.02
BTBG06Q	0.73***	0.01	0.92***	0.03	0.13
BTBG06N	0.70***	0.01	0.79***	0.07	0.09
BTBG06A	0.52***	0.02	0.42**	0.13	0.09
BTBG06D	0.58***	0.01	0.68***	0.09	0.11
Unconstrained Model fit: CFI = 0.949, TLI = 0.910, RMSEA = 0.068, Chi Square = 424.91, df = 17, p 0.001, AIC = 53,359.45, BIC = 53,563.06 Constrained Model fit: CFI = 0.948, TLI = 0.932, RMSEA = 0.059, Chi Square = 438.91, df = 23, p 0.001, AIC = 53,394.42, BIC = 53,558.62					

(continued)

[2] One teacher per school

Table 15.3 (continued)

Factor					
Interaction with colleagues	Within level		Between level		
	Loadings	SE	Loadings	SE	
BTBG09E	0.84***	0.01	0.98****	0.01	0.14
BTBG09C	0.72***	0.01	0.92***	0.03	0.11
BTBG09F	0.79***	0.01	0.89***	0.03	0.18
BTBG09A	0.67***	0.01	0.88***	0.05	0.10
BTBG09G	0.73***	0.01	0.91***	0.03	0.16
BTBG09D	0.63***	0.01	0.73***	0.08	0.29
Unconstrained Model fit: CFI = 0.982, TLI = 0.968, RMSEA = 0.047, Chi Square = 212.58, df = 17, $p < 0.001$, AIC = 55,935.67, BIC = 56,139.28 Constrained Model fit: CFI = 0.980, TLI = 0.972, RMSEA = 0.043, Chi Square = 236.65, df = 22, $p < 0.001$, AIC = 55,936.32, BIC = 56,107.09					
Respect of rules	Within level		Between level		
	Loadings	SE	Loadings	SE	
BTBG07D	0.82***	0.01	0.88***	0.05	0.09
BTBG07E	0.84***	0.01	0.99***	0.02	0.08
BTBG07F	0.84***	0.01	0.91***	0.05	0.09
BTBG07G	0.60***	0.02	0.58***	0.13	0.05
BTBG07H	0.63***	0.02	0.63***	0.11	0.07
BTBG06M	0.56***	0.02	0.78***	0.09	0.09
Unconstrained Model fit: CFI = 0.998, TLI = 0.997, RMSEA = 0.014, Chi Square = 29.47, df = 15, $p < 0.001$, AIC = 49,703.95, BIC = 49,920.70 Constrained Model fit: CFI = 0.994, TLI = 0.991, RMSEA = 0.023, Chi Square = 78.51, df = 21, $p < 0.001$, AIC = 49,754.16.54, BIC = 49,931.50					

***$p < 0.001$. # of observations = 5261, # of clusters = 45

Table 15.4 provides the results of the measurement invariance test. The comparison is between the unconstrained model, as a basis for fit quality, and the constrained models (metric = weak invariance constraints; scalars = strong invariance constraints). In comparison to the *multilevel analysis*, the multiple group analysis is more sensitive to the within variance. This means that the invariance test compares countries to one another and does not provide two-level structures, as in the multilevel model (school and country levels).

Regarding the configural and the metric model, we found similar fit quality. Based on our comparisons of the 45 countries, which generated a large variance, the CFI differences between the metric and the configural models were: 0.031, 0.009, 0.012, and 0.018 for 'teachers' profession', 'caring for learning', 'interaction with colleagues', and 'respect of rules', respectively. These findings demonstrate that the difference was insignificant, except for the first factor—'teachers' profession'. Thus, we can cautiously conclude that a general factor structure exists across all countries.

Table 15.4 Multiple group analysis for the SEC dimensions ($N = 5261$ teachers[3])

	Configural model	Metric model	Scalar model	Metric versus configural	Scalar versus configural	Scalar versus metric
Teacher's profession						
CFI	0.945	0.914	0.624	0.031	0.29	0.593
TLI	0.921	0.902	0.642	0.019	0.26	0.623
RMSEA	0.067	0.075	0.142	−0.008	−0.067	0.15
Chi-Square	2627.36	3574.25	8784.80	953.57	6222.58	5227.80
df	1724	2164	2604			
P	< 0.001	< 0.001	< 0.001			
SRMR	0.066	0.134	0.301			
# of parameters	1741	1301	861			
AIC	89,690.47	89,826.48	94,624.99			
BIC	101,125.49	98,371.54	100,280.11			
Care for students' learning						
CFI	0.948	0.939	0.762	0.009	0.186	0.177
TLI	0.902	0.929	0.799	−0.027	0.103	0.13
RMSEA	0.122	0.104	0.175	0.018	−0.053	−0.071
Chi-Square	989.94	1319.26	3680.35	356.49***	2658.46***	2597.25***
df	360	580	800			
P	< 0.001	< 0.001	< 0.001			
SRMR	0.051	0.185	0.265			
# of parameters	855	635	415			
AIC	49,803.50	49,834.12	52,024.49			
BIC	55,419.20	54,004.85	54,750.25			
Interaction with colleagues						
CFI	0.978	0.966	0.748	0.012	0.23	0.218
TLI	0.959	0.961	0.788	−0.002	0.171	0.173
RMSEA	0.078	0.077	0.179	0.001	−0.101	−0.102
Chi-Square	619.03	981.28	3788.96	360.98***	3270.23***	2832.48***
df	360	580	800			
P	< 0.001	< 0.001	< 0.001			
SRMR	0.034	0.112	0.303			

(continued)

[3] One teacher per school

Table 15.4 (continued)

	Configural model	Metric model	Scalar model	Metric versus configural	Scalar versus configural	Scalar versus metric
# of parameters	855	635	415			
AIC	52,863.34	52,790.88	55,419.76			
BIC	58,479.04	56,961.61	58,145.51			
Respect of rules						
CFI	0.992	0.974	0.859	0.018	0.133	0.115
TLI	0.979	0.964	0.866	0.015	0.113	0.098
RMSEA	0.061	0.079	0.153	−0.018	−0.092	−0.074
Chi-Square	386.37	845.47	2649.41	455.16***	2267.21***	1897.55***
df	270	490	710			
P	< 0.001	< 0.001	< 0.001			
SRMR	0.027	0.177	0.260			
# of parameters	945	725	505			
AIC	46,073.11	46,158.10	47,661.85			
BIC	52,279.94	50,919.96	50,978.73			

***$p < 0.001$. # of observations = 5261, # of countries = 4

In other words, the countries share a similar factor structure in 'care for students' learning', 'interaction with colleagues', and 'respect of rules'. However, for the 'teachers' profession' factor, this cannot be concluded. In contrast to the other three dimensions, we may conclude that 'teachers' profession' is perceived differently by teachers across the 45 countries.

The final measurement model. To confirm the full factor structural model, we ran an integrative measurement model on two-thirds of the sample that included all four factors. Table 15.5 shows all factor loadings in an integrative measurement model. Since all loading values are high and similar to one another, the overall confirmation of the factor structure is strong. This supported undertaking further analyses using these latent factors, in comparison to one factor (Common Method Variance [CMV] test – Podsakoff et al., 2003).

A measure of internal consistency—the composite reliability measure—was added, which replaced the common alpha (Raykov, 1997). The composite reliability is a measure of the latent and the observed variance, in which the loadings represent the latent variance. The composite reliability is the ratio between the squared sum of loadings and the variance of the latent factor (set to a unit variance) over the sum of the latent variance from above and the sum of the variances of the observed items. Our results indicated a high level of reliabilities (composite reliability > 0.70) for all

Table 15.5 Confirmatory factor analysis (measurement model), factor loadings and consistency ($N = 5261$ teachers[4])

Factor	Loadings	SE
Factor 1: Teacher's profession; CR = 0.84		
BTBS17D Adapting my teaching to engage students' interest	0.76***	0.01
BTBS17A Inspiring students to learn science	0.69***	0.01
BTBS17F Assessing student comprehension of science	0.67***	0.01
BTBS17C Providing challenging tasks for the highest achieving students	0.70***	0.01
BTBS17G Improving the understanding of struggling students	0.65***	0.01
BTBS17B Explaining science concepts or principles by doing science Experiments	0.64***	0.01
BTBG14G Encourage students to express their ideas in class	0.43***	0.01
BTBG14D Encourage classroom discussions among students	0.45***	0.01
BTBG14C Ask students to complete challenging exercises that require them to go beyond the instruction	0.43***	0.02
BTBG14F Ask students to decide their own problem solving procedures	0.44***	0.02
BTBG14B Ask students to explain their answers	0.35***	0.02
Factor 2: Care for students' learning; CR = 0.87		
BTBG06P Amount of instructional support provided to teachers by school leadership	0.83***	0.01
BTBG06O Collaboration between school leadership and teachers to plan instruction	0.86***	0.01
BTBG06Q School leadership's support for teachers' professional development	0.75***	0.01
BTBG06N Clarity of the school's educational objectives	0.72***	0.01
BTBG06A Teachers' understanding of the school's curricular goals	0.53***	0.01
BTBG06D Teachers working together to improve student achievement	0.62***	0.01
Factor 3: Interaction with colleagues; CR = 0.89		
BTBG09E Work together to try out new ideas	0.85***	0.01
BTBG09C Share what I have learned about my teaching experiences	0.74***	0.01
BTBG09F Work as a group on implementing the curriculum	0.81***	0.01
BTBG09A Discuss how to teach a particular topic	0.69***	0.01
BTBG09G Work with teachers from other grades to ensure continuity in learning	0.77***	0.01
BTBG09D Visit another classroom to learn more about teaching	0.64***	0.01
CR	0.89	
Factor 4: Respect of rules; CR = 0.87		
BTBG07D The students behave in an orderly manner	0.81***	0.01
BTBG07E The students are respectful of the teachers	0.84***	0.01
BTBG07F The students respect school property	0.84***	0.01
BTBG07G This school has clear rules about student conduct	0.62***	0.01

(continued)

[4] One teacher per school

Table 15.5 (continued)

Factor	Loadings	SE
BTBG07H This school's rules are enforced in a fair and consistent manner	0.65***	0.01
BTBG06M Students' respect for classmates who excel in school	0.59***	0.01
CR	0.87	

Goodness-of-Fit: CFI = 0.934, TLI = 0.924; RMSEA = 0.044, SRMR = 0.050; CR = Composite Reliability

four factors. The model fit was above the acceptance level (e.g., CFI = 0.93, TLI = 0.92).

Table 15.5, like in Table 15.2, illustrates that a few SEC dimensions consist of items related to the different original factors of TIMSS and do not come from one original factor (for example, 'teachers' profession' includes items from the original questions, 14 and 17; 'respect of rules' includes items from the original questions, 6 and 7), or part of the items consist the original factor, and not all the items define the original factor (e.g., 'care for students' learning', 'interaction with colleagues').

Thus, regarding our first and second goals, which explored whether a shared perception of 'school ethical culture' emerged, and to find the meaning of SEC, we found a shared perception of SEC which includes four dimensions. Three of these dimensions ('care for students' learning', 'interaction with colleagues', and 'respect of rules') are highly similar across the 45 countries, while the dimension of the teachers' profession is not as similar. Nevertheless, this dimension also significantly appeared at the teachers' level in the exploratory factor analysis (Table 15.2) and in the confirmatory factor analysis (Table 15.5).

The second step: The relationship between SEC and science achievements
In order to evaluate the relationship between the constructed factors of the teachers' perceptions of SEC and students' scores in science, we used the confirmatory factor analysis (CFA) framework on the full sample ($N = 7890$). However, we undertook a separate regression of the students' scores on each factor ('teachers' profession', 'care for students' learning', 'interaction with colleagues', and 'respect of rules').

Figure 15.1 conceptually illustrates the MIMIC (Multiple Indicators Multiple Causes) model, especially the effect of the latent SEC factors on two-level students' scores. The school level's (level 1) slope can be estimated as a random effect; that is, the slope varies between countries. In this illustration, λ represents loadings, ε represents measurement errors, and β represents the regression slopes from the SEC to the science scores. Note that these parameters appear in both the school level and the country level. However, at the country level, all items appear as latent, as they are not explicitly measured, but are extracted from the model.

Plausible data. We used a procedure that considered the five plausible values (Foy, 2017). The TIMSS data provide students' scores in a plausible value format. There are five imputed values that are substituted for the single score per student. Plausible values are imputations that are meant to avoid a single measurement of

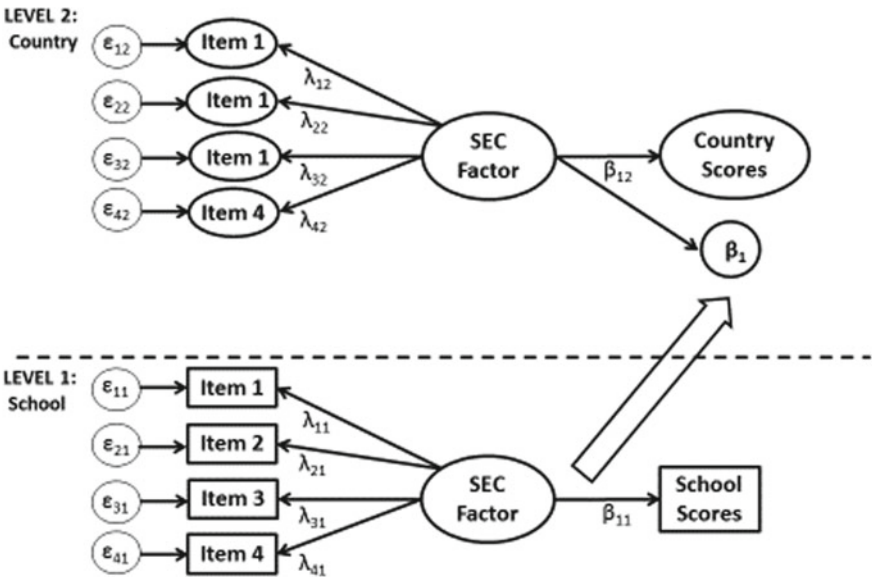


Fig. 15.1 The multilevel model structure for the effect of SEC dimensions on scores in science

a test score, and they include a prior distribution, rather than a point estimate (von Davier, Gonzalez & Mislevy, 2009). This approach is commonly used in large-scale data, which have fewer measurements or one measurement per respondent at level one. Note that for the teachers' data, we retained the plausible procedure by aggregating each plausible value into the school (or teacher's) level. Any further analysis which included scores was then run independently for each value. The mean (across countries) regression slope is reported.

As shown in Table 15.6, the analysis-based MIMIC model in Mplus v.8.0 provided a first answer to the core research question concerning the effect of SEC on students' achievements in science. To run this analysis, we used the multilevel approach for schools at level one and countries at level two (Heck & Ried, 2017). For each factor, the effect on scores in science is presented as unstandardized and standardized regression coefficients. Table 15.6 shows that the factor effects on scores are different in direction (positive/negative) for the school level and for the country level. More specifically, on the school level (within country), all four teachers' SEC dimensions positively affected the mean score of the science exams. The effect was high for respecting rules ($\beta = 0.30, p < 0.001$), moderate for teacher' profession ($\beta = 0.16, p < 0.001$) and caring for learning ($\beta = 0.13, p < 0.001$), and small for interaction with colleagues ($\beta = 0.07$, $p < 0.05$). All effects were significant at $p < 0.05$. However, at the country level (a mean across all schools within the country), we found the opposite relationships: the higher the 'teachers' profession' and the 'interaction with colleagues', the lower the country's mean score in science is ($\beta = -0.47, p < 0.001$; $\beta = -0.40$, $p < 0.01$; respectively). In contrast, in 'care for students' learning', and

Table 15.6 Relationship between the SEC dimensions and science scores ($N = 7890$ teachers[5])

	Unstandard		Standard	
Factor	b	SE	β	SE
Factor 1: Teacher's profession				
Within countries	8.62***	0.97	0.16***	0.02
R2			0.02***	0.005
Between countries	−26.66***	6.81	−0.47***	0.11
R2			0.22***	0.10
CFI = 0.958, TLI = 0.944; RMSEA = 0.030, SRMR = 0.028, Chi-Square = 788.50, df = 98, $p < 0.001$				
Factor 2: Care for students' learning				
Within countries	7.38***	1.37	0.13***	0.02
R2			0.02**	0.006
Between countries	−10.51	8.44	−0.19	0.15
R2			0.03	0.06
CFI = 0.944, TLI = 0.913; RMSEA = 0.056, SRMR = 0.143, Chi-Square = 687.41, df = 27, $p < 0.001$				
Factor 3: Interaction with colleagues				
Within countries	3.89***	1.24	0.07*	0.02
R2			0.005	0.003
Between countries	−22.82**	8.08	−0.40**	0.14
R2			0.16	0.11
CFI = 0.979, TLI = 0.968; RMSEA = 0.037, SRMR = 0.046, Chi-Square = 319.44, df = 27, $p < 0.001$				
Factor 4: Respect of rules				
Within countries	16.47***	1.23	0.30***	0.02
R2			0.09***	0.01
Between countries	7.92	13.92	0.14	0.24
R2			0.02	0.07
CFI = 0.987, TLI = 0.978; RMSEA = 0.029, SRMR = 0.108, Chi-Square = 196.52, df = 25, $p < 0.001$				

*$p < 0.05$. **$p < 0.01$. *** $p < 0.001$

'representing rules', the country level's mean of the teachers' SEC did not have a significant effect on the country mean scores in science.

These findings are illustrated in Figs. 15.2 and 15.3. We found that high achievement countries were low in 'teachers' profession' and 'interaction with colleagues', while low achievement countries were high on these dimensions.

[5] One teacher per school

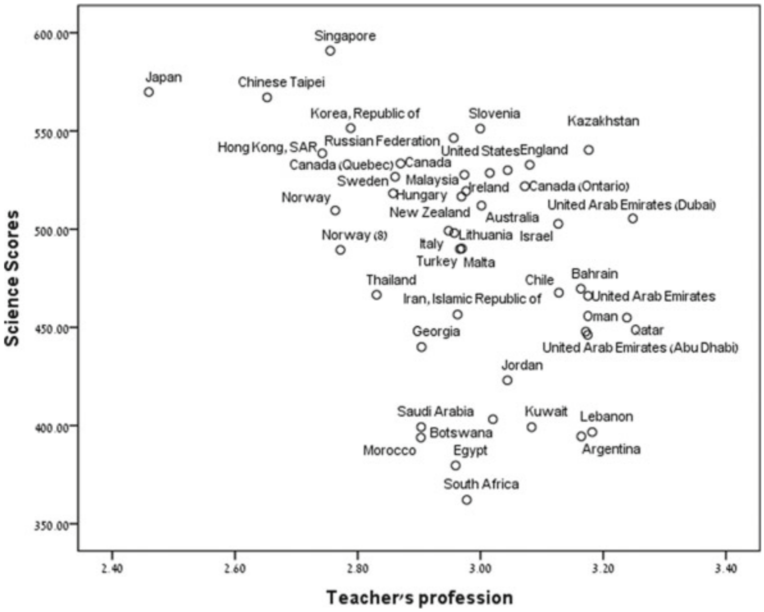


Fig. 15.2 The negative relationship between 'teacher's profession' and science achievements (country level)

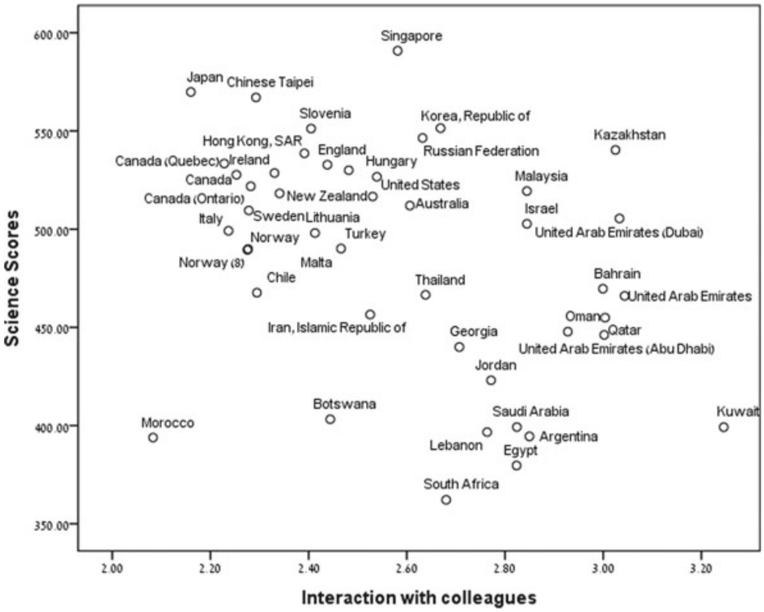


Fig. 15.3 The negative relationship between 'interaction with colleagues' and science achievements (country level)

In sum, regarding our third goal, to investigate the effect of SEC on teachers' perceptions through predicting cross-national science achievements, we conclude that the shared teachers' perceptions of SEC across countries predict students' science achievements. This reflects different levels of effects between the SEC's dimension and the students' science achievements. These differences were also found on the country level.

15.5 Discussion

The main goal of this chapter was to discover whether we could elicit the implicit meaning of the SEC in teachers' TIMSS questionnaires from 45 countries. If we found a shared perception among teachers, we planned to examine whether the SEC, including its dimensions, would predict students' TIMSS eighth grade achievements in science.

This chapter contributes to the literature by examining the meaning of EC among teachers in schools and by eliciting different dimensions of the concept SEC. Specifically, this chapter provides additional meaning to teachers' TIMSS questionnaires by uncovering a new SEC concept agreed upon by experts in ethics in education in different countries. The analysis provided a deeper cross-national meaning for teachers' perceptions of SEC, by exposing the four dimensions of SEC: 'teachers' profession', 'care for students' learning', 'interaction with colleagues' and 'respect of rules'. Furthermore, we found some similarity between our findings and some of the CEV's dimensions ('clarity', 'supportability', 'discussability', and 'sunctionability'). Thus, our finding provides conceptual validity for our generated SEC dimensional model, including the emphasis on the school context.

Specifically, the first dimension, 'teachers' profession', includes teachers' perceptions of their professional standards, such as: adapting teaching, assessing students, and using the method of inquiry, which is similar to the concept of 'clarity' in CEV. This dimension includes perceptions of standards to which teachers are expected to adhere. Our second dimension, 'care for students' learning', includes perceptions of support provided to teachers by the school leadership or school leadership that supports teachers' professional development. This is similar to the concept of 'Supportability' in CEV, which includes aspects of organizational strengthening of the employees. Our third dimension, 'interaction with colleagues', includes perceptions of discussing how to teach and sharing teaching experiences, which are similar to 'discussability' in CEV. This includes aspects of sharing issues, a process in which people learn from each other. Finally, the fourth dimension, 'respect of rules', includes perceptions of rules that are enforced and clear rules about student conduct. This is similar to 'sanctionability' in CEV, which includes perceptions of enforcement, through punishment and rewards.

We understood that universal values may differently affect teachers' perceptions in the participating countries, since these countries differ on aspects, such as culture, context, policy and politics. Therefore, this chapter deepened the sense of universal meaning for teachers' perceptions of SEC, by considering and explaining the relationship between the different dimensions of SEC and students' achievements in the sampled countries.

In particular, the findings about the school level reflected positive relationships between the different dimensions of teachers' perceptions of SEC and students' science achievements. However, on the country level, we found that when the dimensions of 'teachers' profession' and 'interaction with colleagues' were higher, students' achievements were lower. An explanation for this finding might relate to the concept 'marginal addition' (Shapira-Lishchinsky, 2009). In high achievement countries, teachers may perceive the effect of the 'teacher's profession' and 'interaction with colleagues' as not being so important, in comparison to countries that have low academic achievement, since the former are already high on the achievement level. Therefore, these factors do not significantly contribute to their viewpoints. However, in low achievement countries, teachers may attribute significant importance to these dimensions because they may perceive them as having the potential to increase students' achievements.

In sum, this chapter points to a universal perspective, showing how common values, such as different dimensions of SEC, positively affect students' achievements at the school level. However, at the country level, the different impact of these dimensions on student achievement and the different perceptions of 'teachers profession' that were found among the participating countries may be explained by the specific national context which is dependent upon policy and politics regarding these dimensions.

15.6 Conclusions

The findings support a universal perception of SEC among teachers, which was elicited from teachers' responses to TIMSS questionnaires. This led to the understanding that the TIMSS teachers' questionnaire has additional meaning, which goes beyond its original factors. The results also contributed to understanding the meaning of SEC among teachers, by identifying its four dimensions. This may be a new measure that, up until now, has never been investigated in schools. This universal perspective has also supported the ability to predict students' achievements by examining teachers' perceptions concerning their profession and interaction with colleagues. However, the findings also support the importance of national culture. This led to the appearance of different levels of the SEC dimensions, in relation to students' achievements, as well as to differences in the dimension of teachers' profession, which were found to be perceived differently in the countries that were part of this research. The effect of the national culture may be explained by the influence of national policies and norms on school systems, which may affect teachers' perceptions.

Appendix: Selection rule for relevant items (Expert judgment)

The selection rule was as follows: first, count the times each value of relevancy appears across the 10 experts. Second, calculate the product of the relevancy value, multiplied by the times the value appeared in the judgments of the 10 experts. Third, rank the product value from largest to smallest, subject to relevancy value that is equal or greater than '3' (on a ranking scale of 1–5) across all 10 experts.

RANK $\sum_{i=1}^{10} V_I$; Subject to: $V_i > 2$; $i = 1, 2, \ldots, 10$, experts

References

Banks, J. A. (2015). *Cultural diversity and education.* Pearson.

Campbell, E. (2011). Teacher education as a missed opportunity in the professional preparation of ethical practitioners. In L. Bondi (Ed.), *Toward professional wisdom* (pp. 81–88). Ashgate Publishing.

Christians, C. G., Fackler, M., Richardson, K., Kreshel, P., & Woods, R. H. (2015). *Media ethics: Cases and moral reasoning.* Routledge.

Cullen, J. B., Parboteeah, K. P., & Hoegl, M. (2004). Cross-national differences in managers' willingness to justify ethically suspect behaviors: A test of institutional anomie theory. *Academy of Management Journal, 47*(3), 411–421.

Denison, D. R. (1996). What is the difference between organizational culture and organizational climate? A native's point of view on a decade of paradigm wars. *Academy of Management Review, 21*(3), 619–654.

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology, 26*(8), 897–899.

Donnelly, J. (2013). *Universal human rights in theory and practice.* Cornell University Press.

Forsyth, D. R., O'Boyle, E. H., & McDaniel, M. A. (2008). East meets west: A meta-analytic investigation of cultural variations in idealism and relativism. *Journal of Business Ethics, 83*(4), 813–833.

Foy, P. (2017). *TIMSS 2015 user guide for the international database.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and IEA.

Hanushek, E. A., & Woessmann, L. (2015). *The knowledge capital of nations: Education and the economics of growth.* MIT Press.

Heck, R. H., & Reid, T. (2017). Using multilevel regression to examine hierarchical data: Investigating differences in reading performance between immigrant and native-born children. *Culturay Educación, 29*(3), 619–665.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus.* Where was it published? Routledge.

Ho, J. A. (2010). Ethical perception: Are differences between ethnic groups situation dependent? *Business Ethics: A European Review, 19*(2), 154–182.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (Eds.). (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies.* Where was it published? Sage.

Ivison, D. (2010). Justice and imperialism: On the very idea of a universal standard. In missing editors' names, *Law and politics in British colonial thought* (pp. 31–48). Where was it published? Palgrave Macmillan

Kaptein, M. (2011). From inaction to external whistleblowing: The influence of the ethical culture of organisations on employee responses to observed wrongdoing. *Journal of Business Ethics, 98*(3), 513–530.

Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *Journal of Applied Psychology, 15*(1), 1–31.

Klassen, R. M., Usher, E. L., & Bong, M. (2010). Teachers' collective efficacy, job satisfaction, and job stress in cross-cultural context. *The Journal of Experimental Education, 78*(4), 464–486.

Li, S. F., & Persons, O. S. (2011). Cultural effects on business students' ethical decisions: A Chinese versus American comparison. *Journal of Education for Business, 86*(1), 10–16.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data.* John Wiley & Sons.

Melé, D., & Sánchez-Runde, C. (2013). Cultural diversity and universal ethics in a global world. *Journal of Business Ethics, 116*(4), 681–687.

Meyer, M. A., & Booker, J. M. (2001). *Eliciting and analyzing expert judgment: A practical guide.* Where was it published? Society for Industrial and Applied Mathematics.

Milner, H. R. (2010). What does teacher education have to do with teaching? Implications for diversity studies. *Journal of Teacher Education, 61*(1–2), 118–131.

Minkov, M., & Hofstede, G. (2011). The evolution of Hofstede's doctrine. *Cross Cultural Management: An International Journal, 18*(1), 10–20.

Mitonga-Monga, J., & Cilliers, F. (2015). Ethics culture and ethics climate in relation to employee engagement in a developing country setting. *Journal of Psychology in Africa, 25*(3), 242–249.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016a). *TIMSS 2015 international results in mathematics.* TIMSS & PIRLS International Study Center at Boston College.

Mullis, I. V. S., Martin, M. O., & Loveles, T. (2016b). *20 Years of TIMSS: International trends in mathematics and science achievement, curriculum, and instruction.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

O'Boyle, E. H., Forsyth, D. R., & O'Boyle, A. S. (2011). Bad apples or bad barrels: An examination of group and Organizational-level effects in the study of counterproductive work behavior. *Group & Organisation Management, 36*(1), 39–69.

Osborne, J. W. (2015). What is rotating in exploratory factor analysis. *Practical Assessment, Research & Valuation, 20*(2), 1–7.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behaviorial research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.

Rausch, A., Lindquist, T., & Steckel, M. (2014). A test of US versus Germanic European ethical decision-making and perceptions of moral intensity: Could ethics differ within western culture? *Journal of Managerial Issues, 26*(3), 259–283.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173–184.

Riivari, E., & Lämsä, A. M. (2014). Does it pay to be ethical? Examining the relationship between organisations' ethical culture and innovativeness. *Journal of Business Ethics, 124*(1), 1–17.

Ruiz-Palomino, P., & Martínez-Cañas, R. (2014). Ethical culture, ethical intent, and Organizational citizenship behavior: The moderating and mediating role of person–organisation fit. *Journal of Business Ethics, 120*(1), 95–108.

Ruiz-Palomino, P., Martínez-Cañas, R., & Fontrodona, J. (2013). Ethical culture and employee outcomes: The mediating role of person-organisation fit. *Journal of Business Ethics, 116*(1), 173–188.

Sagnak, M. (2010). The relationship between transformational school leadership and ethical climate. *Educational Sciences: Theory and Practice, 10*(2), 1135–1152.

Schein, E. H. (2010). *Organizational culture and leadership* (Vol. 2). John Wiley & Sons.

Scmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210–222.

Shapira-Lishchinsky, O. (2009). Israeli teachers' perceptions of mentoring effectiveness. *International Journal of Educational Management, 23*(5), 390–403.

Shapira-Lishchinsky, O., & Raftar-Ozery, T. (2018). Leadership, absenteeism acceptance, and ethical climate as predictors of teachers' absence and citizenship behaviors. *Educational Management Administration & Leadership, 46*(3), 491–510.

Shiraev, E. B., & Levy, D. (2015). *Cross-cultural psychology: Critical thinking and contemporary applications*. Routledge.

SPSS IBM Corporation Released 2017. *IBM SPSS Statistics for Windows, Version 24.0*. IBM Corporation.

Stromquist, N. P., & Monkman, K. (Eds.). (2014). *Globalization and education: Integration and contestation across cultures*. R & L Education.

Terry, H. (2011). *Golden rules and silver rules of humanity*. Infinite Publishing.

Treviño, L. K., & Weaver, G. R. (2003). *Managing ethics in business organizations: Social scientific perspective*. Stanford University Press.

Treviño, L. K., & Youngblood, S. A. (1990). Bad apples in bad barrels: A causal analysis of ethical decision-making behavior. *Journal of Applied Psychology, 75*(4), 378.

Treviño, L. K., Butterfield, K. D., & McCabe, D. L. (1998). The ethical context in organizations: Influences on employee attitudes and behaviors. *Business Ethics Quarterly, 8*(3), 447–476.

Tullberg, J. (2015). The golden rule of benevolence versus the silver rule of reciprocity. *Journal of Religion and Business Ethics, 3*(1), 223–354.

von Davier M., Gonzalez E. & Mislevy R. (2009). *What are plausible values and why are they useful*? IERI Monograph Series Issues and Methodologies in Large-Scale Assessments. IER Institute. Educational Testing Service.

Wang, J., Lin, E., Spalding, E., Odell, S. J., & Klecka, C. L. (2011). Understanding teacher education in an era of globalization. *Journal of Teacher Education, 62*(2), 115–120.

Zhao, Y. (2010). Preparing globally competent teachers: A new imperative for teacher education. *Journal of Teacher Education, 61*(5), 422–431.

Orly Shapira-Lishchinsky is a Professor and Head of the Department of Educational Administration, Leadership, and Policy at Bar-Ilan University, Israel. Her research fields include educational leadership and management, comparative education, international assessment in science and math (TIMSS), organizational ethics, mentoring by team-based simulation, and teachers' withdrawal behaviors. She was awarded the European Union's Horizon 2020 for her research: "Interdisciplinary model of schooling." Based on her expertise, she serves as an Expert MSCA-IF evaluator. She has published a wide range of highly ranked scholarly journals and a book *Organizational Ethics in Educational Systems*. Lately, she was invited as a keynote speaker to Albany State University of New York, NY; Michigan State University; and the Open University of Cyprus.

Part V
Multilevel Modeling in Educational Research

Chapter 16
Why They Want to Leave? A Three-Level Hierarchical Linear Modeling Analysis of Teacher Turnover Intention

Lixia Qin

Abstract Researchers from various disciplines are developing more complex understandings of phenomena by using multilevel lenses. In an effort to explore the behaviors and practices of individuals, groups, schools and even countries, educational researchers should expand both theoretical and empirical investigations to encompass these multilevel considerations. This chapter is an effort to draw more policy attention to multilevel studies in the field of teaching force that might provide a response to some debates with regard to teacher distribution and turnover found in single-country and single-level studies. The multilevel analysis in this chapter underscores the joint impact and the interactive effects of individual and situational factors on teacher turnover and the variation across 32 countries is a function of teacher-, school- and country-level factors. This chapter has demonstrated how teaching force research can significantly benefit from using multilevel analysis to more explicitly investigate the macro/micro effects on teachers than the existing studies have been able to do. The implication for stressing multilevel lenses to promote educational development will also be discussed.

Keywords HLM model · Multilevel research · Teacher turnover intention · Cross-national analysis · Country contexts

16.1 Introduction

Research on teacher turnover focuses heavily on individual and school factors in explaining turnover decisions (e.g. Hanushek & Rivkin, 2007; Ingersoll, 2001). Single-country studies, however, have revealed mixed results. For example, whereas some studies suggest that teacher salary is an influential factor in teachers' career decisions (Goldhaber et al., 2007), others indicate the poor predictability of teacher payment (Hanushek & Rivkin, 2007).

L. Qin (✉)
University of Wisconsin System, Madison, WI, USA
e-mail: lqin@uwsa.edu

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_16

Research restricted to the within-country perspective tends to overlook country effects on teacher turnover. For example, theory suggests that opportunity wages outside of the teaching field should have less of an influence on turnover intentions in countries where teaching is a high-status job than in countries where teaching lacks prestige (Falch & Strøm, 2005). This study argues that whether or not a teacher decides to transfer to another school or quit teaching altogether is not determined solely by his or her own individual characteristics or the school where he or she works. Teacher choice also needs to be investigated through a focus on both macro and micro effects and micro/macro interactions. As this study will show, teacher turnover intention is the outcome of multilevel effects. Even though microlevel factors play a crucial role in predicting teacher turnover, macrolevel effects also shape turnover intention, and there are significant cross-level interaction effects. The findings suggest that some of the teacher and school effects might be conditional on the country-level context. For example, the analysis finds that some individual and school variables have a stronger effect on turnover intention in countries where the status of the teaching profession is higher.

This study employs multilevel statistical models to understand cross-national variation in teacher turnover intention. Survey data from the Organization for Economic Co-operation and Development (OECD) are utilized to describe teacher transfer intention and teaching attachment across countries. The OECD has become the authoritative, international source for comparative information about educational outcomes, policies and practices (Sellar & Lingard, 2013). The primary data come from the 2013 Teaching and Learning International Survey (TALIS) conducted by the OECD. The study merges the TALIS surveys with other international data on national context to gain insight into, and a better understanding of, factors contributing to teacher turnover intentions within lower secondary schools (grades 7–9) across countries. This chapter has two main objectives: first, to investigate the direct effects of country variables on teacher turnover intention by controlling for the compositional effects from lower-level factors (teacher and school characteristics) and second, to analyze the moderation effects of country variables on outcomes via cross-level interaction analysis.

16.2 Multilevel Modeling and Teacher Turnover Research

Moving from a single level of analysis (e.g., individual, group/team, organization, and country), the increasing research across fields has been emphasizing on the multilevel perspective. Multilevel theories explain individual or group attributes and outcomes within the multiple contexts. Multilevel research examines one or more systems or environments that are most often hierarchically nested within one another (e.g., peer groups nested within schools) (Dunn et al., 2014). Multilevel approach recognizes the integration of individual and system and how they interactively shape individual and organizational outcomes (Kozlowski & Klein, 2000).

Growing body of literature has focused on developing the theoretical perspectives and methodologies to incorporate a multilevel analysis of institutions, organizations and individual level (Fligstein, 2001; Sillince et al., 2001). Researchers across fields, such as organization (Cooney, 2007), social psychology (Dunn et al., 2014) and Human Resource (Upton & Egan, 2010), have discussed the potential and challenges of employing multilevel theoretical framework in research. In the field of management, for example, in the illustration of new understanding of corporate social change activities, Bies et al. (2007) suggested that "involves examination of corporate social agency at multiple levels of analysis: the micro level (focusing on psychological and social psychological bases), the meso level (involving relational and network issues), and the macro level (involving political, economic, institutional and societal dynamics)" (p. 789).

Researchers from education are developing more complex understandings of phenomena by using multilevel lenses. A multilevel lens may help us reveal the complexity and richness of social behavior and "it draws our attention to the context in which behavior occurs and illuminates the multiple consequences of behavior traversing levels of social organization" (Hitt et al., 2007, p. 232). For researchers in education continuously seeking to explain the behaviors and practices of individuals, groups, schools and even countries, it is important to expand educational theories and empirical investigations to encompass these multilevel effects.

Conventional employee turnover models are built on the individual's rational choices and assume that the individual considers available employment options in order to maximize personal benefit (Minniti & Lévesque, 2008). Those models, however, fail to capture the social forces that may drive or constrain career decisions (Haltiwanger et al., 2014). Moreover, a multilevel framework is important to a cross-national study such as this because solely macrolevel analyses based on aggregated data may result in issues such as omitted variables and measurement error (Haltiwanger et al., 2014).

Research on teacher turnover has centered around two separate thematic strands. One strand of research focuses on school characteristics. This body of literature emphasizes the impact of managerial practices and school climate on teacher perceptions of working conditions and their teaching performance at school level of analysis (e.g., Ingersoll, 2001). On the other hand, researchers who are interested in studying teacher turnover and mobility at the individual level of analysis (e.g., Boyd et al., 2005) have linked teachers' individual characteristics to their labor market decisions. Both approaches have made significant contributions to explaining teacher turnover and mobility. However, neither approach can sufficiently explore and account for this aspect. The school-level and/or district-level only approach may overlook the meaningful individual differences, while the individual-level only approach may ignore the contextual factors that can significantly shape and constrain individual teachers' job decisions.

Solely examining one level may fail to understand teacher turnover in a more comprehensive perspective and some crucial factors may be overlooked. In addition, the interaction effects between different levels may also be disregarded due to the one-level approach. The factors across different levels can interact with each other and jointly determine teacher turnover (Kozlowski & Klein, 2000). In the case of teacher labor market, the multilevel consideration can be helpful in clarifying the causes of teacher attrition. The multilevel approach will be more sufficient and accurate in estimating the combined effect of individual teacher characteristics and the larger unit attributes on turnover intention.

Additionally, according to the multilevel theories, a top-down approach assesses organizational factors (contextual factors) that can affect individual perceptions and behaviors. A bottom-up method can reduce the variability of individual perceptions and behaviors by explaining the emergence of collective phenomena (Kozlowski & Klein, 2000). Therefore, in this chapter, the relevant contextual features will also be examined and hypothesize that high-level factors would have top-down influences on teacher turnover intention via both a direct and a moderating effect. Furthermore, individual teachers' attributes and perceptions would combine to form a collective phenomenon at school level through bottom-up processes and would significantly relate to teacher labor market at schools.

Furthermore, it is suggested that study should be explicit about how data at lower level can be related to another level (Chen et al., 2005). Researchers propose two principle in terms of data aggregation for identifying the relationship between lower-level data and higher-constructs (Hitt et al., 2007). The first is composition. It refers to the use of simple data description to interpret how to connect lower-level data with higher-level constructs. For example, in the current study, some perceptions of teachers at the first level will be aggregated into the school level to represent school organizational characteristics. Aggregated attitudes and perceptions have been seen as one of the important categories of organizational characteristics, such as the quality of leadership and management, and organizational culture and climate (Hausknecht & Trevor, 2011). The second principle is complication. In this principle, the lower-level data will be integrated in a more complex and nonlinear way (Kozlowski & Klein, 2000). The current study has individual teachers answer survey items about their own work attitudes and perceptions about their school. Therefore, those responses would be naturally aligned with individual constructs at the individual level. Meanwhile, the study also plans to use the individual-level data to test hypotheses about school phenomena.

16.3 Teacher Turnover Intention

In this study, the phrase "turnover intentions" refers to teachers' attitude favoring leaving their current workplace or profession (Tiplic et al., 2015). Studying turnover intentions can help to forecast the actual turnover (e.g., Steel & Ovalle, 1984). According to some psychological studies, intent is a strong predictor of behavior (e.g., Lee & Whitford, 2008; Price & Mueller, 1981). Even though the relationship between turnover intentions and actual turnover may vary across studies, for instance, this relationship can be moderated by variables such as labor market conditions (Kirschenbaum & Weisberg, 1990) or, some suggest it may also depend on the employee's motivation (Vandenberghe & Tremblay, 2008). Recent evidence shows that there is consistent evidence indicating that turnover intentions are a strong precursor of actual turnover behaviors (Cho & Lewis, 2012). In addition, another rationale of using turnover intention as a turnover proxy is its accessibility and the desirable statistical qualities (Cohen et al., 2016), and it's more economic (Dalton et al., 1999).

The chapter has separately examined the different groups in turnover intention: the teachers with transfer intention and teachers with quit intention. According to Ingersoll (2001), movers are the teachers who transfer or move to another school, but still stay in teaching profession. Leavers refer to teachers who quit their teaching job altogether. Teacher turnover, in much of the empirical research, has only been focused on those who leave their teaching position, whereas teachers who move to another school or even district have been understudied. Although there are some studies on teacher mobility, less attention has been paid on it because teacher mobility in general doesn't increase or decrease the overall number of teachers, also both of them have been viewed to have same effect on schools (Ingersoll, 2001).

The studies separately examining the transfer and exit attrition reveal that different factors may have impacted on teachers' job decisions: quit or move to other schools. For example, teachers who exit school system tend to be more sensitive to salary changes (Gritz & Theobald, 1996). Some findings have shown that the teachers who exit are relatively more competitive than and those who still stay in teaching profession, and they are more likely to have a competitive education background (Boyd et al., 2005). Therefore, separating leaver and mover can have valuable policy implications. For instance, if the differences between them are small, policymakers can have more confidence in state-wide policies that affect all districts in similar ways (Imazeki, 2005). On the other hand, if the characteristics of transfers are significantly different from leavers within a district or even state, then policies regarding teacher recruitment and retention may need to be more directed and specific so as to have more targeted response. For example, some districts may have higher proportion of teachers transfer to other schools and some districts may have more teachers leave teaching position (Imazeki, 2005).

Studying teacher turnover intention is also very important for identifying the "reluctant stayer". As contextual factors have been reported as important determinants to teacher turnover intention (e.g., Moynihan & Pandey, 2008), some social and administrative factors in some countries may limit teachers' alternative job opportunities. Thus, even if a teacher who is dissatisfied intends to leave or quit, he/she may still choose to stay and keep the job, which means the actual turnover will not be observed but the issue remains, as often happens in developed countries. Research proposes the needs of identifying this group because the reluctant stayers often appear as "bad apples" (Felps et al., 2006). The effect of reluctant stayers might be worse because low job satisfaction and high stress have been found to negatively influence teachers' work enthusiasm and decrease productivity work, which certainly will impact on students' learning and development (Sargent & Hannum, 2005).

By predicting turnover intentions administrators could provide targeted retention strategies to teachers at risk of leaving (Boyd et al., 2011), and specific support to those "reluctant stayers" who feel *trapped* and disengaged in their schools (Li et al., 2016).

16.4 Teacher- and School-Level Characteristics and Teacher Turnover Intention

Although the compositional effect is not this study's emphasis, the models include a series of teacher- and school-level variables that capture lower-level characteristics to control for the compositional differences across countries.

Scholars in numerous countries have identified a variety of reasons why teachers transfer to a different school or leave the teaching profession. Those reasons can be categorized mainly into teacher and school attributes. Teacher characteristics include teacher demographics, teaching experience and education. For example, consistent empirical findings have revealed that attrition is more common among young teachers (Hanushek et al., 2004; Ingersoll, 2001) and novice teachers are more likely to leave the profession in the early stages of their career (Ingersoll & Smith, 2003; Tiplic et al., 2015). The literature on gender differences in teacher turnover shows mixed results. Some scholars find that female teachers are more likely to quit than their male counterparts (Gritz & Theobald, 1996), while others observe the opposite (Ingersoll, 2003). In addition, teacher education also contributes to the variance in teacher turnover. Teachers with more extensive teacher education backgrounds tend to persist in the teaching field (Ahn, 2015; Lankford et al., 2002).

Components relating to school attributes are identified in the research as student characteristics (e.g., Bonhomme et al., 2016; Hanushek et al., 2005), school size and class size (Brill & McCartney, 2008), school location (Feng, 2014), and student disciplines (e.g., Borman & Dowling, 2006). Teacher turnover rates tend to be significantly higher in schools serving disadvantaged students (Bonhomme et al., 2016; Hanushek et al., 2004). Similar findings have been reported in countries such as Sweden (Karbownik, 2016) and Norway (Falch & Strøm, 2005). A school's geographic location also has been found to impact teachers' choices. For example, teachers tend to leave urban schools for suburban districts (Feng, 2014). School size is also associated with teacher turnover. Some findings reported higher attrition in large, urban schools (e.g., Brill & McCartney, 2008; Lankford et al., 2002). A considerable amount of research has demonstrated that smaller schools provide a more collegial environment and are less likely to lose teachers (e.g., Newmann & Wehlage, 1995). In addition, student discipline is one of the most-cited reasons for teachers' decisions to quit (e.g., Borman & Dowling, 2006; Brill & McCartney, 2008). The issue is even more significant among beginning teachers, who say they experience more pressure regarding their relationship with students and their ability to manage student behavior (e.g., Luekens et al., 2004). The study has also included working hours and the teacher-student ratio as school attributes. Research suggests that teachers' positive sense of their status was closely shaped by their working conditions. Working hours and teacher-student ratio are two of the most important factors shaping teachers' working conditions (Hargreaves & Flutter, 2013).

16.5 Country-Level Variables and Teacher Turnover Intentions

While research on teacher distribution and turnover focuses heavily on individual teacher characteristics (e.g., experience, education, age) (e.g., Boyd et al., 2008; Whipp & Geronime, 2017), more recent work has expanded the research to school organizational characteristics that may affect teachers' decisions to leave their schools (e.g., Falch & Strøm, 2005; Newton et al., 2018). Limited work, however, has analyzed teacher turnover as an individual teacher decision nested within larger social contexts (Yang et al., 2018).

The considerable differences have been observed across countries regarding the relationship between teacher and school and teacher turnover due to numerous factors (e.g., the structure of the teacher labor market and the policy effort of the governments) (OECD, 2005). A variety of nation- and region-specific rules are making the teacher labor market different from private sectors and also various across countries, for example, wage schedule and job promotion scale, and the teacher personnel policies. In addition, the levels of centralization of the education system can also influence the teacher labor market and the extent teachers' career choices based on their own preferences (OECD, 2005).

Most studies of teacher turnover build on within-country analysis because of the advantages of single-country research, such as the fairly constant institutional settings, and the very limited comparable data across countries (Agasisti & Zoido, 2015). Yet, solely focusing on one country can lead to insularity. Cross-countries comparison may gain more valuable insights and opportunities regarding policies, and those policies can never be transferred without considering the differences between countries. The analyses of national trends by using aggregate data may show us little about the realities of individual schools, which are crucial for understanding of the situation at the national or regional levels. The perceptions of teachers and principle may, on the other hand, more or less demonstrate the international trends and how they are being experienced at a country or local level.

While education policy makers around the world have paid increasing attention to attracting and retaining high-quality teachers (OECD, 2014; UNESCO Institute for Statistics, 2006), many countries struggle with teacher shortages and high turnover rates (OECD, 2005, 2014). Forces beyond the school context, such as the low social status of the teaching profession, relatively low salaries, and increasing opportunities for alternative jobs, have been highlighted across countries (UNESCO Institute for

Statistics, 2006). Under this multilevel framework, along with individual and school factors, country-specific effects also can influence the antecedents of teacher turnover. The study hypothesizes that cross-national differences in teacher status and alternative employment opportunities would be significantly related to between-country differences in teacher turnover intention.

Teaching Status

The first country context in this study is teaching status. The author hypothesizes that the cross-national variation in the teaching status can explain the differences in teacher career decisions. Many teachers around the world feel that their work is undervalued. The social status of teachers in some East Asian countries, such as Japan and South Korea, is relatively high (Kim & Han, 2002). In the societies where the teaching profession is highly valued, students seem to be more academically successful (Burns & Darling-Hammond, 2014), and the teacher workforce is usually more stable and more likely to attract highly qualified graduates (OECD, 2014). In contrast, teachers' commitment to their job decreases in countries where teaching is a low-status profession (Symeonidis, 2015). To measure how teachers perceive their teaching status, the teachers' responses to the question, "I think that the teaching profession is valued in society" have been selected for the analysis (TALIS, 2013).

Teacher Salary

Another important indicator associated with teachers' career decisions is teacher salary. Empirical evidence across countries suggests that the significant variations in teacher pay not only are reflected in educational outcomes but also impact teachers' job satisfaction and attrition (Imazeki, 2005; Stockard & Lehman, 2004). The negative correlation between salaries and teacher turnover has been identified across the scholarly literature (e.g., Hanushek et al., 2004; Hendricks, 2014; Ingersoll, 2001).

The available international evidence shows that teacher pay has declined over the last 30 years and has not kept up with salaries of other occupations in some countries, especially in low-income countries (e.g., Leigh & Ryan, 2008). Findings from the Teacher Status Index indicate that respondents in most of the countries considered their teachers to be underpaid (TSI, 2014). A study of teacher salaries from 1999 to 2013 demonstrated a significant cross-country difference regarding changes in the relative earnings of teachers (Varga, 2017). This study focused on whether the variation of teacher salaries accounts for the differences of teacher turnover intention across countries. This study has used the relative salary information offered by OECD (Education at a Glance, 2014). The relative salary indicator is calculated based on teachers' salaries relative to earnings for full-time, full-year workers with tertiary education in each country. The indicator has been adjusted for inflation using the deflators for private consumption. The data showed that teachers' relative salaries varied significantly across countries. Korea, Portugal, and Spain have highest relative salaries. Teacher salaries in those countries are at least 20% higher than those of workers with tertiary education. The Czech Republic and Slovak Republic have the lowest relative salaries (on average, less than 50% of those of workers with a tertiary education) (Education at a Glance, 2015).

Furthermore, research has shown that besides relative salary, the range of teacher salary increases also has a significant impact on teacher turnover (Imazeki, 2005; Varga, 2017). In OECD countries, the salary at the top of the scale (after teachers reach around 24 years of experience) increased by 64% over starting salaries, on average. However, the between-country variation is significant. For example, in some countries (e.g., Denmark and Iceland), the ratio of salary at the top of scale to starting salary is less than 25%, whereas in Luxembourg and Korea, the difference is an average of 80% (Education at a Glance, 2011). In order to see whether a larger salary increase would retain teachers, the second salary variable for this study was the ratio of salary at the top of the scale to starting salary. This indicator also comes from the Education at Glance administrated by OECD in 2014.

General Economic Conditions

The second country context that may have an influence on teacher turnover is the country's general economic conditions. Sound economic conditions may offer job opportunities or alternative labor market opportunities for teachers and are linked to increased teacher turnover and a decline in teacher quality (Roberts et al., 2005). Scholars have noted that the overall academic aptitude of teachers has declined relative to other workers with college degrees in recent decades due to increased opportunities in other fields (Hoxby & Leigh, 2004; Leigh & Ryan, 2008). Unemployment rates in each country can be used to represent the conditions of its labor market. Research findings show that turnover rates of workers in countries with low unemployment rates are expected to be higher than in a country where jobs are scarce (Chew et al., 2016; Hulin et al., 1985).

Due to the limited number of alternative job opportunities in the public sector, however, it may not be sufficient to use a country's total unemployment rate as a predictor (Steel & Griffeth, 1989). Another indicator of alternative employment opportunities in the study was how people perceive outside job opportunities. The indicator came from the Human Development Index (2015), which specifically asked participants from each country how they perceived their local labor market. The value of the indicator refers to the percentage of respondents answering "good" to the Gallup World Poll question, "Thinking about the job situation in the city or area where you live today, would you say that it is now a good time or a bad time to find a job?" Even though workers' perceptions of alternative job opportunities do not necessarily align with actual labor market conditions, it still might influence their turnover intentions (Hwang & Kuo, 2006).

Additionally, per capita GDP and educational expenditures as a percentage of GDP for each country were adopted as control variables in order to more accurately capture the effect of teacher pay on turnover intention. The level of economic development, measured as per capita GDP, usually will influence government expenditures on education (Busemeyer, 2007). Researchers have found a positive link between spending per student and per capita GDP (Hanushek & Luque, 2003). Furthermore, a relationship has been found between government expenditures on education and the level of teacher pay (Glewwe et al., 2011). By controlling for these variables, the study assessed whether teachers in countries with higher salaries are more likely to stay in the profession.

16.5.1 Cross-Level Interactions (Moderation Effects)

In addition to direct effect, national contexts may influence teachers' work attitude and turnover behavior indirectly through individual and school characteristics and practices (Luschei & Chudgar, 2017). For the indirect effect of country variables, the study has mainly focused on the extent to which country variables impact relationships between school disadvantages and teacher turnover intentions. While the bulk of evidence from various countries suggests that teachers tend to leave schools that have high proportions of low-income and minority students, whether the relationship between student disadvantage and teacher turnover intention varies across countries remains unknown.

One important indicator that has been used in the cross-level interaction analysis is wage decisions. Under rigid wage settings, the variation in teacher salaries may not be large enough to compensate for teaching in unattractive schools and neighborhoods (Falch & Strøm, 2005; Feng, 2014) and it is challenging for wages to quickly respond to teacher supply and demand and job attributes (Boyd et al., 2003). TALIS surveyed principle who had significant responsibility for establishing teachers' starting salaries, including setting pay scales. The results showed that in some countries, teacher pay was largely decided at the school level (e.g., the Czech Republic, England, Estonia, the Netherlands, and Sweden), whereas in many other countries, salary decisions are made at the state level or by central administrators. It is hypothesized that working conditions should have more of an influence on turnover intentions in countries where wages are set by higher-level authorities than in countries where wage differentials can compensate for local characteristics such as school-level salary decisions.

Much of the teacher turnover research has focused only on those who leave teaching, whereas teachers who move to another school or district have been understudied since in general, this shift does not affect the overall number of teachers (Grissom et al., 2016). As Ingersoll (2001) noted, "movers" are those who transfer to another school but still stay in the teaching profession and "leavers" quit teaching altogether. Studies that independently examined teacher turnover revealed the factors that impact teacher decisions to transfer or quit are not necessarily the same (Kukla-Acevedo, 2009). Thus, distinguishing between leavers and movers can have policy implications. For instance, if the differences between them are small, policymakers can have more confidence in state-wide policies that affect all districts in similar ways (Imazeki, 2005). On the other hand, if the characteristics of movers differ significantly from leavers, then policies regarding teacher recruitment and retention may need to be more directed and specific to generate a more targeted response (Imazeki, 2005). Due to the availability of data, instead of teachers' intention to quit, the current study has separately examined teachers' intention to transfer and quit intention (as measured through teacher responses to the question, “If I could decide again, I would still choose to work as a teacher”).

16.6 Method

The findings of this study have been primarily built on TALIS2013, the largest international survey of teachers and principle. The database differs from other well-known international data (e.g., PISA, TIMMS) on its focus on the working conditions of teachers and the learning environment in schools. The data permit a detailed description of teacher and principle demographics and school and organizational characteristics and provide robust, policy-relevant indicators (OECD, 2014). The total sample includes 104,358 teachers in 6,455 schools across 32 countries and economies. A set of teacher and school characteristics was identified as lower-level independent variables. Meanwhile, a set of country-level measures that potentially related to the teacher labor market in general and teacher turnover in particular also were included.

Under the multilevel conceptual framework, three main effects have been included to explore the factors relating to teacher turnover intention across countries. The first is the compositional effect that specifies that cross-country differences arise from the unequal distribution of lower-level characteristics (teacher and school factors in this study). In other words, if individual and school characteristics explain, to some extent, a teacher’s turnover intention and if these characteristics vary across nations, then they also can explain the cross-country differences in turnover intention. The second is the contextual effect that occurs when national variables (e.g., teachers’ relative salaries, unemployment rates, teaching status) directly contribute to the differences in teacher turnover intention across countries. The multilevel framework helps us determine whether the country differences in teacher turnover intentions are due to the characteristics of the individuals who live in these countries (compositional effects) or due to factors that relate to the countries themselves (contextual effects). The third is cross-level interaction effect. This occurs when national variables impact the relationship between lower-level characteristics and outcome. In the current study, the cross-level effect refers to the degree to which teacher and school factors influence turnover intentions conditional on country-level context (Ruiter & Van Tubergen, 2009). Namely, this analyzes how country variables may moderate the relationship between lower-level factors and outcomes.

Figure 16.1 provides the study’s conceptual framework. The model consists of three main conditions, with individual variables shown as level one; school-specific variables as level two; and country variables as level three. The solid arrows reflect the fixed effects of predictors at levels 1, 2 and 3 on the outcome. The dotted arrows represent predictors of slopes as outcomes and reflect cross-level moderate effects, which can maximize the potential of hierarchical, linear modeling. The analysis tests the joint effect of individual and school-level variables (compositional effects) and country-level variables (contextual and cross-level interaction effects) on teacher turnover intentions.

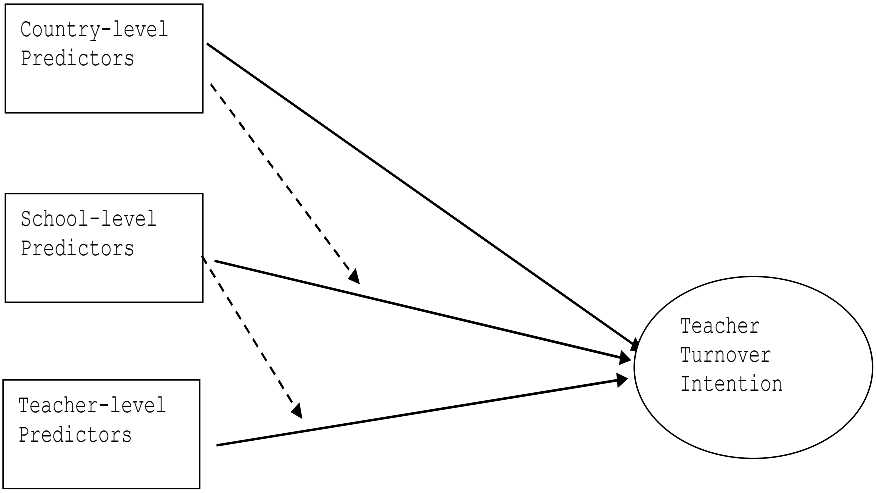


Fig. 16.1 Conceptual model examining the factors relating to teacher turnover intentions across countries

A three-level, hierarchical linear modeling was used to analyze the extent to which differences in outcomes reflect the effects of country, school and individual-specific features. In the first step, a null model was built for both transfer intention and quit intention to establish a baseline model from which subsequent models could be compared, and also to capture the intraclass correlation coefficient (ICC). The second step was to build intercept-and-slope-as-outcome models to capture both lower and higher-level effects on the outcomes and to test the cross-level interaction effects (moderation effects) of country variables on teacher turnover intention. Two different models, quit intention model and transfer model, have been built to capture different aspects of turnover intention. For each model, the study has focused on the effects of country variables: teaching status (perceived teaching social status, working hours, teacher-student ratio and teacher salaries), alternative job opportunities (satisfaction with the local labor market, unemployment rate) to teachers' turnover intentions.

16.7 Findings

Before examining the teacher turnover intention across countries, the study first looked at the issue of teacher shortages across countries through the data. There was a large inter-country variation in principle' reports regarding the impact of teacher shortages/ teacher inadequacies (see Fig. 16.2). On average, 38% of principle believed that shortages/teacher inadequacies were an issue in their schools, ranging from 13% in Poland to 80% in Japan. While principle' responses cannot easily be taken at face

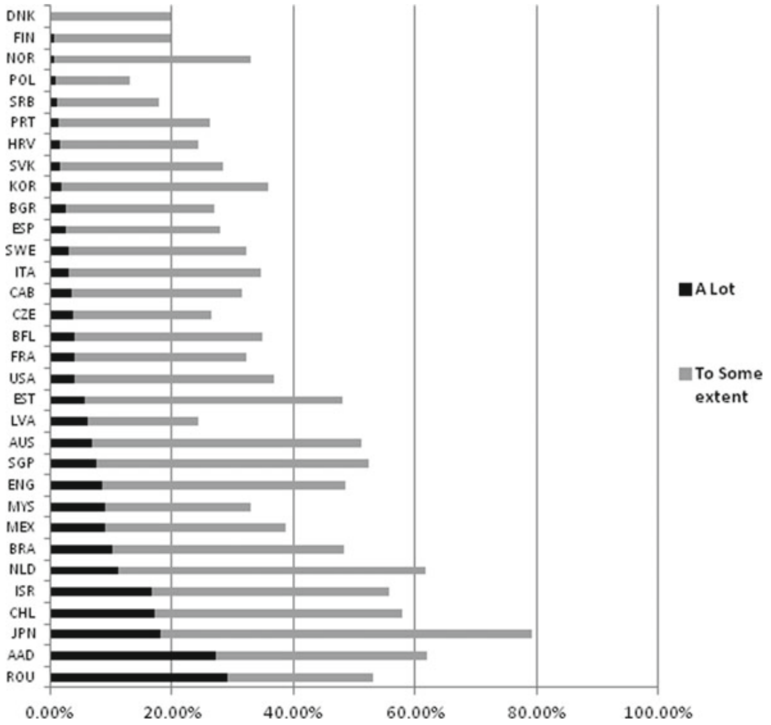


Fig. 16.2 National average of principle who agreed that the shortage of qualified teachers in their school had hindered their capacity to provide quality instruction (as percentage)

value, they can be viewed as perceptions of whether teacher shortages existed in their schools.

Figure 16.3 illustrates a large between-country variation in both transfer and quit intention of the teachers. For the transfer intention ($M = 21.3$, $SD = 8.01$), most countries have a relatively large proportion of teachers who tend to move to other schools. For example, over 40% of teachers in Malaysia expressed an intent to transfer. The right side of the chart indicates the percentage of teachers who did not agree with the statement: “I regret that I decided to become a teacher” (Quit Intention). Again, there was a large variation across countries regarding teachers’ quit intention ($M = 30$, $SD = 9.36$). At 10%, Malaysia had the smallest proportion, while Sweden, with 47%, had the largest.

The first step of the HLM analysis was to create an unconditional model to partition the total variance in the outcome variable into each level of the data. Level-1, level-2 and level-3 unconditional models, which did not include any predictors at any level, were developed. The results suggested that significant variation existed among teachers within schools, across schools within countries and across countries in both models. The intraclass correlation (ICC), which represents the proportion of the variance in the transfer intention model, was 0.085 and 0.11 at the country and

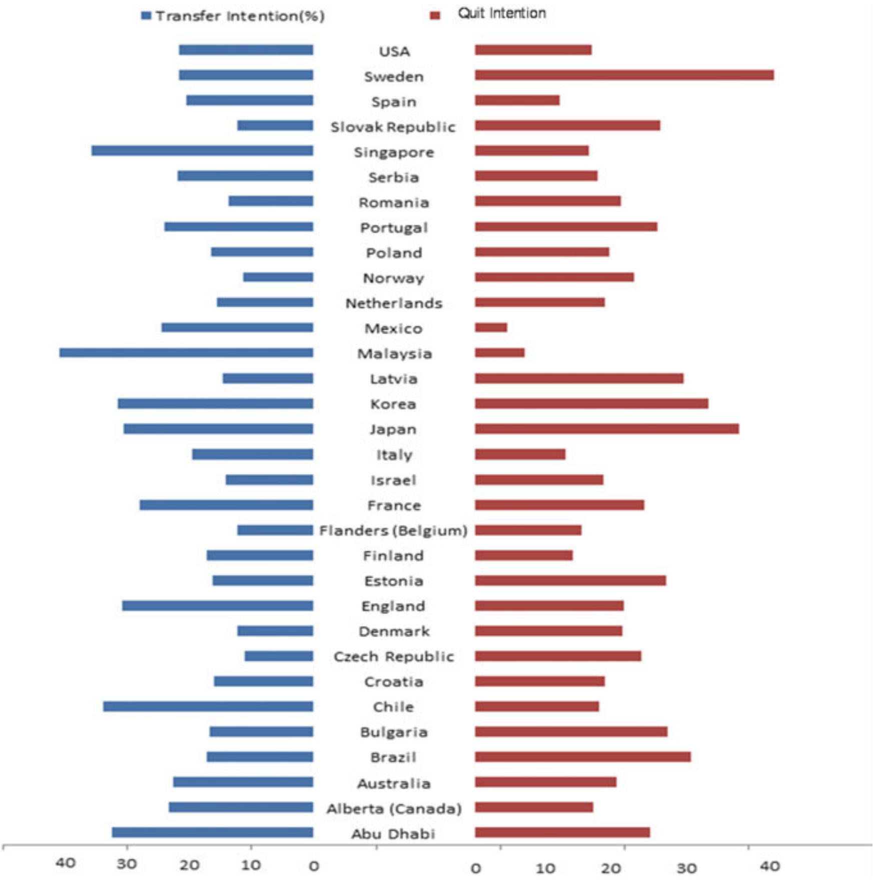


Fig. 16.3 Turnover intention and teaching attachment

school level. This shows that 8.5% and 11% of the total variance in transfer intention was accounted for by country- and school-level differences, respectively. The rest of the variance 80.5% [1 − (0.085 + 0.11)] was due to within-school differences. In the quit intention model, the ICC values at the country and school level were 0.087 and 0.12. This shows that 8.7% and 12% of the total variance in quit intention was accounted for by country- and school-level differences, respectively. The rest of the variance of 79.3% [1 − (0.087 + 0.12)] was due to within-school differences. Even though some ICC values were relatively small, the multilevel models utilized for them still had a substantial impact on the inferences.

16.7.1 The Effect of Individual and School Characteristics (Compositional Effects)

As shown in Table 16.1, individual characteristics captured a substantial portion of cross-country variance in teacher turnover intentions. For example, regarding demographic variables, younger, male teachers were significantly more likely to consider changing schools (Table 16.1). Compared with female teachers, male teachers were more likely to intend to leave their workplace or teaching position altogether. Teachers who taught math and those with higher educational attainment showed a significantly higher intention to quit. Teachers with higher educational attainment also showed higher intention to transfer to another school. For job characteristics, classroom discipline proved significantly positively related to turnover intention ($p < 0.001$). The teachers who reported more classroom discipline issues were more likely to intend to transfer or quit. Working hours had a negative impact on transfer intention.

School size also had a significant impact on transfer intention. As school size increased, teacher intention to transfer decreased. The percentage of low-income students in school was positively related to teachers' transfer intention. Rural teachers were more likely to consider switching schools. Still, these associations did not hold across all models. Working hours were negatively related to teacher intention to leave ($r = -0.006$, $p < 0.05$). This correlation, again, was no longer significant in the model considering teacher salary. The teacher-student ratio had a different effect on teacher detachment and transfer intention. This variable was positively associated with teacher transfer intention in the model ($r = 0.06$, p < 0.05). Teachers in countries with larger teacher-student ratios were more likely to change schools. Similarly, teacher-student ratio was positively associated with quit intention ($r = 0.02$, $p < 0.01$). Quit intention was higher in countries with higher teacher-student ratios and the significance remained after adding salary variables.

16.7.2 The Effects of Country Variables

16.7.2.1 The Direct Effects

In the transfer intention models, the variable of perceived teaching status was an important contextual predictor (see Table 16.1). The positive correlation indicated that the more teachers believed that society valued their job, the more likely they were to switch schools ($r = 0.26$, $p < 0.01$). Still, this effect was no more significant after adding salary variables ($N = 21$). Relative salary has a positive effect on teacher transfer intention ($r = 0.76$, $p < 0.01$). Teachers from countries with higher relative salaries tended to change schools more than those from countries with lower relative salaries.

Table 16.1 Three-level effects on teachers' transfer intention and quit intention

Fixed effects	Transfer intention model (N = 32)	Transfer intention model with salary (N = 21)	Transfer intention model without salary (N = 21)	Quit intention (N = 32)	Quit intention with salary (N = 21)	Quit intention model without salary (N = 21)
INTRCPT	1.89*** (0.04)	1.91*** (0.027)	1.91*** (0.027)	1.85*** (0.042)	1.87*** (0.013)	1.87*** (0.027)
Gender	0.036*** (0.007)	0.045*** (0.009)	0.033*** (0.007)	0.07*** (0.006)	0.09*** (0.009)	0.05*** (0.009)
Age	−0.002** (0.006)	−0.0018* (0.007)	−0.002** (0.0015)	−0.0006 (0.002)	0.0006 (0.0007)	−0.0018* (0.007)
Math	−0.037* (0.008)	−0.048* (0.011)	−0.036* (0.014)	−0.035** (0.008)	−0.033** (0.011)	−0.032* (0.011)
Science	0.01 (0.007)	0.009 (0.01)	0.02 (0.014)	0.001 (0.009)	0.06*** (0.01)	0.009 (0.01)
Working hours	−0.0009** (0.0002)	−0.0007* (0.0003)	−0.0006* (0.0004)	−0.0005* (0.000)	−0.00003 (0.0003)	−0.0007* (0.0003)
Education	0.056*** (0.008)	0.068*** (0.013)	0.053*** (0.01)	0.005*** (0.008)	0.04** (0.013)	0.006*** (0.013)
Experience	−0.004** (0.0007)	−0.003** (0.0007)	−0.003** (0.0001)	−0.003** (0.0008)	0.002** (0.0007)	−0.004** (0.0007)
% of low-income in classroom	0.034*** (0.004)	0.012** (0.007)	0.027*** (0.007)	0.02*** (0.005)	0.0004 (0.0007)	0.02 (0.007)
% of students with behavioral issues in classroom	0.031** (0.004)	0.071*** (0.008)	0.034** (0.008)	0.02** (0.002)	0.064*** (0.009)	0.03** (0.008)
School predictors						
School size	−0.00007** (0.000)	−0.00006** (0.000)	−0.00006** (0.000)	0.00005 (0.006)	0.00002 (0.000)	0.00006 (0.000)
% of ELL	0.009 (0.007)	0.005 (0.01)	−0.007 (0.01)	−0.009 (0.006)	−0.015 (0.008)	0.005 (0.01)
% of special ed	−0.02 (0.01)	0.01 (0.01)	−0.01 (0.08)	0.005 (0.007)	−0.0024 (0.008)	0.01 (0.01)
% of low SES	0.023** (0.005)	0.054*** (0.01)	0.022** (0.003)	0.02** (0.02)	0.013 (0.007)	0.02** (0.01)
Rural	0.048** (0.02)	−0.01 (0.01)	−0.041** (0.02)	−0.026 (0.01)	−0.006 (0.015)	−0.02 (0.01)
Urban	−0.02 (0.013)	0.01 (0.01)	−0.01 (0.016)	0.005 (0.01)	−0.017 (0.01)	0.001 (0.01)
Working hours	0.002 (0.004)	−0.009 (0.004)	0.004 (0.006)	−0.006* (0.005)	0.0085 (0.013)	0.007* (0.004)

(continued)

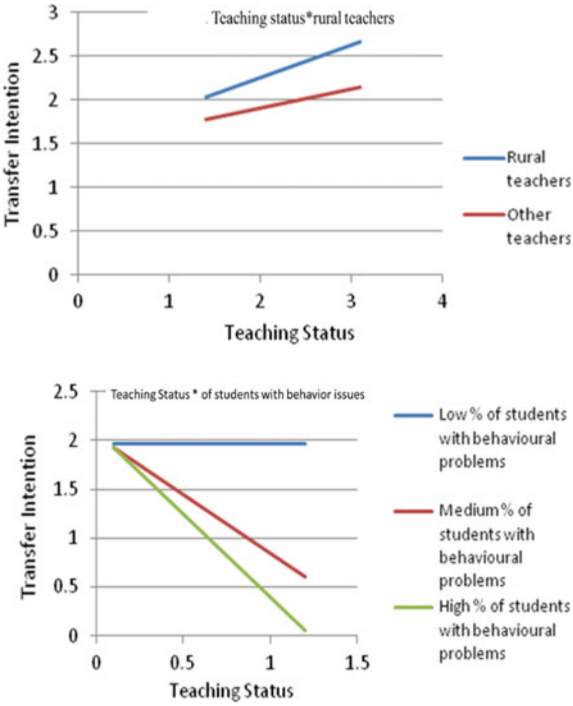
Table 16.1 (continued)

Fixed effects	Transfer intention model (N = 32)	Transfer intention model with salary (N = 21)	Transfer intention model without salary (N = 21)	Quit intention (N = 32)	Quit intention with salary (N = 21)	Quit intention model without salary (N = 21)
Teacher-student ratio	0.06* (0.007)	0.036** (0.008)	0.02* (0.007)	0.03** (0.008)	0.035** (0.005)	0.036** (0.008)
Country predictors						
Perceived teacher status	0.26* (0.08)	−0.004 (0.084)	0.23* (0.07)	−0.274*** (0.038)	−0.435*** (0.05)	−0.291*** (0.084)
Labor market sentiment (%)	0.0021 (0.0042)	0.0025 (0.0018)	0.0023 (0.0018)	0.004 (0.006)	0.001 (0.001)	0.002 (0.002)
Unemployment rate	−0.006 (0.003)	−0.032 (0.09)	−0.032 (0.09)	−0.025* (0.006)	0.002 (0.006)	−0.032* (0.09)
Relative salary		0.763* (0.1556)			−0.646** (0.163)	
Salary increase		−0.136 (0.155)			0.225 (0.1)	
Wage decision	0.009 (0.001)	0.006 (0.003)	0.008 (0.002)	0.002 (0.001)	0.003 (0.002)	0.002 (0.001)
GDP	−0.00003 (0.000)	0.00003 (0.000)	0.00002 (0.000)	0.00001 (0.000)	−0.00006 (0.000)	0.00003 (0.000)
Government expenditures (% of GDP)	−0.055* (0.015)	−0.102* (0.017)	−0.061* (0.027)	0.058 (0.02)	0.03 (0.001)	0.045 (0.027)
Model deviance (parameters)	183,166.2 (108)	176,952.41 (110)	182,888.6 (108)	181,566.5 (108)	140,361.86 (110)	181,952.41 (018)

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$

In the quit intention models (Table 16.1), perceived teacher status was a strong predictor for quit intention ($r = -0.27$, $p < 0.001$). The result showed that teachers had lower levels of quit intention in countries where they believed teaching was prestigious (also see Fig. 16.4). In contrast to the transfer intention model, relative salaries had a negative effect on quit intention ($r = -0.65$, $p < 0.01$). Teachers in countries where teachers were paid well *in comparison with* other college graduates were less likely to consider leaving.

Fig. 16.4 The cross-level interactions



16.7.3 *The Cross-Level Interaction (The Moderation Effect of Country Variables)*

One of the purposes of this study was to assess the moderation effect of country-level factors on the relationships between the lower-level factors and outcomes. Cross-level interactions are useful for answering questions about why lower-level effects vary across higher-level units (Raudenbush & Bryk, 2002). The study has focused on the extent to which the teacher- and school-level effects varied across countries, with particular attention paid to whether country variables may alter the relationship between student disadvantage and teacher turnover intention. The analysis revealed some significant cross-level interactions for both the transfer and quit intention model (see Table 16.2).

Transfer Intention Model

As Table 16.2 indicates, as teaching status increased, the effect of rural location on teacher transfer intention increased. That is, rural teachers in countries with higher levels of teaching status were more likely to consider changing schools ($r = 0.13$, $p < 0.01$) (see Fig. 16.4). In contrast, as teaching status increased, the effect of a high proportion of students with behavioral issues on teachers' transfer intention

Table 16.2 The cross-level interactions

Transfer intention	
Teaching status * rural	0.1291 (0.046) **
Teaching status * % of students with behavior issues	−0.05 (0.075)*
Salary increase * education	−0.144 (0.054)**
Relative salary * experience	0.008 (0.004)*
Unemployment rates	0.132 (0.05)**
Wage decision * working hours	0.0005 (0.003)*
Quit intention	
Teaching status * education	−0.06 (0.001)*
Salary increase * SES	0.067 (0.021)**
Wage decision * teacher-student ratio	−0.0002 (0.001)**

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$

decreased. This means that teachers with a high percentage of students with behavioral problems were less likely to change schools in high-teaching-status countries than those in low-teaching-status countries ($r = -0.05$, $p < 0.01$) (see Fig. 16.4). Salary increases weakened the correlation between education degree and transfer intention ($r = -0.14, p < 0.01$). Teachers with higher educational attainment were less likely to change schools in countries with larger salary increases (starting/maximum teachers' statutory salaries). Those with less teaching experience were more likely to change schools in countries with higher relative salaries. Unemployment rates had a positive effect on rural teachers' transfer intention. As the unemployment rate increased, rural teachers' intent to change schools increased, as well. Wage decisions had a positive effect on the correlation between working hours and transfer intention. As the percentage of state-level salary decisions increased, the effect of working hours on teachers' transfer intention increased, as well.

Quit Intention Model

The strength of the correlation between educational attainment and quit intention decreased in countries where teaching status was high. Teachers with higher educational attainment were less likely to leave their job in countries with higher teacher status ($r = -0.06$, $p < 0.05$). Salary increases strengthened the relationship between school socioeconomic status and quit intention. In countries with larger salary increases, teachers from low socioeconomic schools were more likely to consider quitting their job. Wage decisions weakened the relationship between teacher-student ratio and quit intention. The teacher-student ratio has less effect on the outcome in countries with high percentages of state-level wage decision-making.

16.8 Conclusion and Implications

This chapter provides in-depth discussion of how country contexts along with teacher and school variables might relate to teachers' turnover intentions by using a set of three-level HLM models. Using a large sample of teachers and schools from 32 OECD countries, the study estimates a set of three-level HLM models of turnover intention. The multi-country data sets offer information in terms of individual, school, and country effects, respectively, along with the interactions between them (cross-level effects). It is conceptualized that the drivers of teacher turnover intention have a multilevel structure. The differences in outcomes reflect the differences in the effects of country-specific features and the characteristics of the school and individual. Moreover, a multilevel modeling approach helps us split the variance in teacher turnover into three levels: teacher, school and country.

This study focuses on how country differences in teacher turnover intentions are explained by multilevel effects. Multilevel approach can bridge the individual and higher-level perspectives and provide a more comprehensive picture of the teacher characteristics and contextual characteristics (school, districts and even countries) that may contribute to teachers' labor market decisions. Moreover, the interactions across levels have revealed the impact of lower-level attributes on teacher turnover differed in various country contexts. Also, the multilevel perspective is helpful in examining to what extent teacher characteristics and perceptions of working conditions, when aggregated to the school and higher level, could explain between-school and/or between-country differences in observed teacher turnover intention.

The findings confirmed that not only attributes of individuals and schools, but also macro conditions, matter. The variation in teacher turnover intention across 32 countries was a function of teacher-, school-, and country-level factors. The multilevel approach emphasizes the variability that exists at the individuals as well as environment level (Kozlowski & Klein, 2000). Thus, individuals within environments and environments embedded in larger contexts can vary according to one or more dimensions. At the collective level, some researchers call for the context-specific investigations of turnover and the recognition that contextual factors can shape the influence of turnover's antecedents (Hausknecht & Trevor, 2011; Nyberg & Ployhart, 2013).

The findings have captured the extent of country variables attributed to cross-country variance in teacher turnover intention by controlling for compositional effects. Perceived teaching status was one of the most important predictors in the study and was significant across almost all of the models (with or without a consideration of salaries). Teachers' quit intention was lower in countries where teachers thought their profession was respected and valued ($p < 0.000$). Meanwhile, teachers were more likely to switch schools in countries where teaching had a high-social status. One explanation could be that teachers from countries with high-teaching status have more autonomy and freedom/confidence in choosing where they want to teach.

Consistent with previous work, the results showed that salaries can explain, to some extent, the cross-country differences of teacher turnover intentions (Imazeki, 2005). Teachers' relative salaries had a negative effect on quit intention, meaning that teachers in countries with higher relative salaries tended to stay in education. That is, teachers in countries with higher relative salaries were more likely to change schools. It is difficult for a within-country study to obtain an effective measure of salary-to-teacher turnover since most public school teachers in the same country are paid very similarly (Dolton & Marcenaro-Gutierrez, 2011). This multilevel study underscores the important role of teacher salaries in teachers' career decisions by showing that teachers in countries that invested more in teacher salaries reported lower levels of quit intention. In addition, it is clear that once taking into account the effects of teacher salaries, some other country variables (e.g., working hours in the quit intention model and perceived teaching status in the transfer intention model) did not make a difference to the outcome. It also reflects the influential role of teacher salaries in their career decisions.

Another advantage of conducting multilevel approach is to detect some institutional variations that may not be captured through single-country study. Specifically, it can reveal how effects systematically vary across different settings (Hanushek & Woessmann, 2017). This study has indicated that besides the direct effect, the national contexts have influenced teachers' work attitude and turnover behavior indirectly through school practice. For instance, educational system and teacher policy might affect the level of school autonomy, which in turn may influence on teacher labor market decisions (Luschei & Chudgar, 2017). Thus, multilevel analysis could be useful in the investigation of institutional variation that is hard to be fully observed within a country. It is also important to take into account cross-level effects while studying country effects on teacher turnover intentions. In addition to the direct effects, country variables may also have moderation effects on teacher turnover intention. This study found that some important predictors (e.g., student disadvantage) had differential effects on teacher turnover intention under different country contexts. For example, the relationship between the proportion of students with behavioral issues and teacher transfer intention varied as a function of the country-level variable. That is, teachers with a high percentage of students with behavioral problems had lower levels of transfer intention in the countries where teaching had a higher status.

Significant cross-level interaction between wage decisions and working hours may have echoed previous research findings suggesting that rigid teacher salary schedules make differences across working conditions for teachers more substantial (Hanushek & Rivkin, 2007). For example, research has shown teachers are 12 percentage points more likely to be dissatisfied with long working hours than other graduates (Chevalier et al., 2004).

As a result, challenging schools tend to face more severe teacher shortages and retain less-qualified teachers (Bonhomme et al., 2016). With small differences in average pay, improving teacher workloads and/or school working environments might become a more important factor in enticing teachers to stay.

Teaching is one of the most challenging professions even though it is of lower status than many other professions (e.g., medicine, law, and engineering) (Liu & Onwuegbuzie, 2014; National Education Association of the United States, 2003). People around the world choose to teach for a variety of reasons, but all teachers need to be recognized and respected for their profession (MacBeath, 2012). The findings of this study have stressed the role of government in promoting a positive image of teachers and raising public awareness of the value of the teaching profession (Bushaw & Lopez, 2011). The study revealed that teacher salaries and working conditions are not the only important factors in teacher retention; the ability of countries to successfully recruit and retain quality teachers also depended on the status of teaching. Nevertheless, teaching status is a complex concept and contains multiple aspects (Bushaw & Lopez, 2011). Various factors that involve the profession (e.g., social and economic development, characteristics of education systems, school organization) need to be considered in order to effectively and comprehensively improve teaching status.

Increasingly countries have introduced financial incentives into teacher salary structure while designing teacher retention policies (Dolton & Marcenaro-Gutierrez, 2011). However, despite of the importance of teacher pay, numerous studies have addressed concerns over the lack of effective and optimal pay packages in securing high-quality teachers, and a pay raise itself may not be efficient enough to improve teacher retention and teaching performance (Hanushek & Rivkin, 2007). This multilevel study has also indicated that teacher turnover intention is influenced not just by how much teachers have been paid, but also by their working conditions. For example, the finding showed that the ratio of students to teachers that directly reflects on class size was an important predictor of both teachers' transfer intention and quit intention, even after controlling for salary information. A low teacher-student ratio will make teachers' work even more demanding and frustrating at a level that is not offset by high pay (Carnoy & DeAngelis, 2002).

Turnover research has moved from "one size fits all" to contextual-based factors that are more or less important to turnover decisions in a given setting (Hom et al., 2017). This study simultaneously examined individual, school and national contextual factors to provide an intriguing picture by specifying a conceptual framework for cross-level phenomena. The findings underscore the joint impact and the interactive effects of individual and situational factors relating to teachers' turnover intention.

This study contributes to the growing international research on teacher turnover pointing out the necessity of simultaneously assessing the effects of both micro and macro levels on teacher turnover across nations. The findings advance the teacher turnover literature in two ways. First, it addresses the lack of cross-national studies in teacher turnover that explicitly connect micro and macro levels of analysis. While not perfect, this study provides a more comprehensive picture of how country contexts may impact teacher turnover intention. This contribution ties directly to the call for an increase in turnover research to better capture context (Hausknecht & Trevor, 2011). Researchers have warned that over-emphasis on intra-national studies may cause insularity that potentially could lead to insensitivity concerning teacher policies in various situations (Dolton & Marcenaro-Gutierrez, 2011).

Second, the study advanced the empirical literature through its test of cross-level interactions. The findings shed light on some lower-level factors that may have a differential effect on teacher turnover intentions in different country contexts. For example, some within-country research has repeatedly shown that teachers in disadvantaged schools are more likely to quit or move to other schools (Brill & McCartney, 2008; Lankford et al., 2002). Less is known on how these relations vary across countries. This research contends that the relationship between student disadvantage and teacher turnover in some contexts is less significant than in other contexts. A country where teaching is a high-status profession, for example, may have played a role in weakening or even breaking such a correlation. Teachers from countries where teaching is valued seem to be more willing to stay, regardless of the status of school disadvantage.

16.8.1 Limitations

This study has some limitations. First, all the factors were self-reported by teachers and principle. The possible method or respondent bias should not be ruled out. And the reliability of the findings is limited to the reliability of the data sources used in the study: international surveys and government reports. Second, because this was a correlational study based on a cross-sectional dataset, any cause and effect implications are not guaranteed. Third, the variance across countries was still significant, strongly calling for variables to enhance the explanatory power of the models. Although we focused on several country-level variables, other unknown (omitted) factors may have contributed to this unexplained variance. There were no data on other intermediate levels such as school districts. It might be possible that the effects of the omitted levels were reflected in the individual-level estimates. Fourth, the small number of countries in the salary model ($N = 21$) may cause potential sampling bias.

References

Agasisti, T., & Zoido, P. (2015). *The efficiency of secondary schools in an international perspective: Preliminary results from PISA 2012* (OECD Education Working Papers).

Ahn, T. (2015). Matching strategies of teachers and schools in general equilibrium. *IZA Journal of Labor Economics, 4*(1), 1.

Bies, R. J., Bartunek, J. M., Fort, T. L., & Zald, M. N. (2007). Introduction to special topic forum: Corporations as social change agents: Individual, interpersonal, institutional, and environmental dynamics. *Academy of Management Review, 32*, 788–793.

Bonhomme, S., Jolivet, G., & Leuven, E. (2016). School characteristics and teacher turnover: Assessing the role of preferences and opportunities. *The Economic Journal, 126*, 1342–1371.

Borman, G. D., & Dowling, N. M. (2006). The longitudinal achievement effects of multi-year summer school: Evidence from the Teach Baltimore randomized field trial. *Educational Evaluation and Policy Analysis, 28*, 25–48.

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The role of teacher quality in retention and hiring: Using applications to transfer to uncover preferences of teachers and schools. *Journal of Policy Analysis and Management, 30*(1), 88–110.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Who leaves? Teacher attrition and student achievement* (No. w14022). National Bureau of Economic Research.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2003). *Analyzing the determinants of the matching public school teachers to jobs: Estimating compensating differentials in imperfect labor markets* (NBER Working Papers 9878).

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). Explaining the short careers of high-achieving teachers in schools with low-performing students. *American Economic Review, 95*(2), 166–171.

Brill, S., & McCartney, A. (2008). Stopping the revolving door: Increasing teacher retention. *Politics and Policy, 36*(5), 750–774.

Burns, D., & Darling-Hammond, L. (2014). *Teaching around the world: What can TALIS tell us*. Stanford Center for Opportunity Policy in Education. https://edpolicy.stanford.edu/sites/default/files/publications/teaching-around-world-what-can-talis-tell-us_3.pdf

Busemeyer, M. R. (2007). Determinants of public education spending in 21 OECD democracies, 1980–2001. *Journal of European Public Policy, 14*(4), 582–610.

Bushaw, W. J., & Lopez, S. J. (2011). Betting on teachers: The 43rd annual Phi Delta Kappa/Gallup Poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 93*(1), 9–26.

Carnoy, M., & DeAngelis, K. (2002). *The teaching workforce: Concerns and policy challenges*. Organisation for Economic Co-operation and Development, ed. Education Policy Analysis.

Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods, 8*(4), 375–409.

Chevalier, A., Harmon, C., Walker, I., & Zhu, Y. (2004). Does education raise productivity, or just reflect it? *The Economic Journal, 114*(499), F499–F517.

Chew, H. G., Ng, K. Y. N., & Fan, S. (2016). Effects of alternative opportunities and compensation on turnover intention of Singapore PMET. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, 10*(3), 720–728.

Cho, Y. J., & Lewis, G. B. (2012). Turnover intention and turnover behavior: Implications for retaining federal employees. *Review of Public Personnel Administration, 32*(1), 4–23.

Cohen, G., Blake, R. S., & Goodman, D. (2016). Does turnover intention matter? Evaluating the usefulness of turnover intention rate as a predictor of actual turnover rate. *Review of Public Personnel Administration, 36*(3), 240–263.

Cooney, K. (2007). Fields, organizations, and agency: Toward a multilevel theory of institutionalization in action. *Administration &amp; Society, 39*(6), 687–718.

Dalton, D. R., Johnson, J. L., & Daily, C. M. (1999). D.R. Dalton, J.L. Johnson, C.M. Daily On the use of "intent to…" variables in organizational research: An empirical and cautionary assessment. *Human Relations, 52*, 1337–1350.

Dolton, P., & Marcenaro-Gutierrez, O. D. (2011). If you pay peanuts do you get monkeys? A cross-country analysis of teacher pay and pupil performance. *Economic Policy, 26*(65), 5–55.

Dunn, E. C., Masyn, K. E., Yudron, M., Jones, S. M., & Subramanian, S. V. (2014). Translating multilevel theory into multilevel research: Challenges and opportunities for understanding the social determinants of psychiatric disorders. *Social Psychiatry and Psychiatric Epidemiology, 49*(6), 859–872.

Falch, T., & Strøm, B. (2005). Teacher turnover and non-pecuniary factors. *Economics of Education Review, 24*(6), 611–631.

Felps, W., Mitchell, T. R., & Byington, E. (2006). How, when, and why bad apples spoil the barrel: Negative group members and dysfunctional groups. *Research in Organizational Behavior, 27*, 175–222.

Feng, L. (2014). Teacher placement, mobility, and occupational choices after teaching. *Education Economics, 22*(1), 24–47.

Fligstein, N. (2001). Social skill and the theory of fields. *Sociological Theory, 19*(2), 105–125.

Glewwe, P. W., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2011). *School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010* (NBER Working Paper 17554).

Goldhaber, D., Gross, B., & Player, D. (2007). *Are public schools really losing their best? Assessing the career transitions of teachers and their implications for the quality of the teacher workforce* (Working Paper 12). National Center for Analysis of Longitudinal Data in Education Research.

Grissom, J. A., Viano, S. L., & Selin, J. L. (2016). Understanding employee turnover in the public sector: Insights from research on teacher mobility. *Public Administration Review, 76*(2), 241–251.

Gritz, R. M., & Theobald, N. D. (1996). The effects of school district spending priorities on length of stay in teaching. *Journal of Human Resources, 31*(3), 477–512.

Haltiwanger, J., Scarpetta, S., & Schweiger, H. (2014). Cross country differences in job reallocation: The role of industry, firm size and regulations. *Labour Economics, 26*, 11–25.

Hanushek, E. A., & Luque, J. A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review, 22*(5), 481–502.

Hanushek, E. A., & Rivkin, S. G. (2007). Pay, working conditions, and teacher quality. *The Future of Children, 17*, 69–86.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (NBER Working Paper w11154).

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326–354.

Hanushek, E. A., & Woessmann, L. (2017). School resources and student achievement: A review of cross-country economic research. In *Cognitive abilities and educational outcomes* (pp. 149–171). Springer.

Hargreaves, L. & Flutter, J. (2013). *The status of teachers and the teaching profession: A desk-study for education international* (Unpublished manuscript). Department of Education, University of Cambridge, UK.

Hausknecht, J. P., & Trevor, C. O. (2011). Collective turnover at the group, unit, and organizational levels: Evidence, issues, and implications. *Journal of Management, 37*(1), 352–388.

Hendricks, M. D. (2014). Does it pay to pay teachers more? Evidence from Texas. *Journal of Public Economics, 109*(1), 50–63.

Hitt, M. A., Beamish, P. W., Jackson, S. E., & Mathieu, J. E. (2007). Building theoretical and empirical bridges across levels: Multilevel research in management. *Academy of Management Journal, 50*(6), 1385–1399.

Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology, 102*(3), 530.

Hoxby, C. M., & Leigh, A. (2004). Pulled away or pushed out? Explaining the decline of teacher aptitude in the united states. *The American Economic Review, 94*(2), 236–240.

Hulin, C. L., Roznowski, M., & Hachiya, D. (1985). Alternative opportunities and withdrawal decisions: Empirical and theoretical discrepancies and an integration. *Psychological Bulletin, 97*(2), 233.

Hwang, I., & Kuo, J. (2006). Effects of job satisfaction and perceived alternative employment opportunities on turnover intention: An examination of public sector organizations. *Journal of American Academy of Business, 8*(2), 254–259.

Imazeki, J. (2005). Teacher salaries and teacher attrition. *Economics of Education Review, 24*(4), 431–449.

Ingersoll, R. M. (2001). *Teacher turnover and teacher shortages: An organizational analysis* (Working Paper 9-1-01). Graduate School of Education, University of Pennsylvania, Philadelphia.

Ingersoll, R. M. (2003). Turnover and shortages among science and mathematics teachers in the United States. In *Science teacher retention: Mentoring and renewal* (pp. 1–12). NSTA Press.

Ingersoll, R. M., & Smith, T. M. (2003). The wrong solution to the teacher shortage. *Educational Leadership, 60*(8), 30–33.

Karbownik, K. (2016). The effects of student composition on teacher turnover: Evidence from an admission reform. *Economics of Education Review, 75*, 101960.

Kim, E., & Han, Y. (2002). *Attracting, developing and retaining effective teachers: Background report for Korea.* Korean Educational Development Institute.

Kirschenbaum, A., & Weisberg, J. (1990). Predicting worker turnover: An assessment of intent on actual separations. *Human Relations, 43*(9), 829–847.

Kozlowski, S. W., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In *Multilevel theory, research, and methods in organizations.* Wiley

Kukla-Acevedo, S. (2009). Leavers, movers, and stayers: The role of workplace conditions in teacher mobility decisions. *The Journal of Educational Research, 102*(6), 443–452.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37–62.

Lee, S. Y., & Whitford, A. B. (2008). Exit, voice, loyalty, and pay: Evidence from the public workforce. *Journal of Public Administration Research and Theory, 18*(4), 647–671.

Leigh, A., & Ryan, C. (2008). Estimating returns to education using different natural experiment techniques. *Economics of Education Review, 27*(2), 149–160.

Li, J. J., Lee, T. W., Mitchell, T. R., Hom, P. W., & Griffeth, R. W. (2016). The effects of proximal withdrawal states on job attitudes, job searching, intent to leave, and employee turnover. *Journal of Applied Psychology, 101*(10), 1436.

Liu, S., & Onwuegbuzie, A. J. (2014). Teachers' motivation for entering the teaching profession and their job satisfaction: A cross-cultural comparison of china and other countries. *Learning Environments Research, 17*(1), 75–94.

Luekens, M. T., Lyter, D. M., & Fox, E. E. (2004). *Teacher attrition and mobility: Results from the teacher follow-up survey, 2000-01.* National Center for Education Statistics. http://www.nces.ed.gov/pubs2004/2004301.pdf

Luschei, T. F., & Chudgar, A. (2017). Conceptual framework: Marginalized children and their teachers. In *Teacher distribution in developing countries* (pp. 13–23). Springer.

MacBeath, J. (2012). *Future of teaching profession.* Education International.

Minniti, M., & Lévesque, M. (2008). Recent developments in the economics of entrepreneurship. *Journal of Business Venturing, 23*(6), 603–612.

Moynihan, D. P., & Pandey, S. K. (2008). The ties that bind: Social networks, person-organization value fit, and turnover intention. *Journal of Public Administration Research and Theory, 18*(2), 205–227.

National Education Association of the United States. Research Division. (2003). *Status of the American public-school teacher*. National Education Association of the United States, Research Division.

Newmann, F. M., & Wehlage, G. G. (1995). *Successful school restructuring: A report to the public and educators.* Center on Organization and Restructuring of Schools, University of Wisconsin.

Newton, X., Rivero, R., Fuller, B., & Dauter, L. (2018). Teacher turnover in organizational context: Staffing stability in Los Angeles charter, magnet, and regular public schools. *Teachers College Record, 120*(3), 1–36.

Nyberg, A. J., & Ployhart, R. E. (2013). Context-emergent turnover (CET) theory: A theory of collective turnover. *Academy of Management Review, 38*(1), 109–131.

OECD. (2005). *Teachers matter: Attracting, developing and retaining effective teachers.* Organization for Economic Co-operation and Development. https://www.oecd.org/education/school/34990905.pdf

OECD. (2014). *New insights from TALIS 2013: Teaching and learning in primary and upper secondary education.* https://doi.org/10.1787/9789264226319-en

Price, J. L., & Mueller, C. W. (1981). A causal model of turnover for nurses. *Academy of Management Journal, 24*(3), 543–565.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Roberts, L. W., Clifton, R. A., & Ferguson, B. (2005). *Recent social trends in Canada, 1960–2000* (Vol. 12). McGill-Queen's Press-MQUP.

Ruiter, S., & Van Tubergen, F. (2009). Religious attendance in cross-national perspective: A multilevel analysis of 60 countries. *American Journal of Sociology, 115*(3), 863–895.

Sargent, T., & Hannum, E. (2005). Keeping teachers happy: Job satisfaction among primary school teachers in rural northwest China. *Comparative Education Review, 49*(2), 173–204.

Sellar, S., & Lingard, B. (2013). The OECD and global governance in education. *Journal of Education Policy, 28*(5), 710–725.

Sillince, J. A., Harindranath, G., & Harvey, C. E. (2001). Getting acceptance that radically new working practices are required: Institutionalization of arguments about change within a healthcare organization. *Human Relations, 54*(11), 1421–1454.

Steel, R. P., & Griffeth, R. W. (1989). The elusive relationship between perceived employment opportunity and turnover behavior: A Methodological or conceptual artifact? *Journal of Applied Psychology, 74*(6), 846–854.

Steel, R. P., & Ovalle, N. K. (1984). A review and meta-analysis of research on the relationship between behavioral intentions and employee turnover. *Journal of Applied Psychology, 69*(4), 673.

Stockard, J., & Lehman, M. B. (2004). Influences on the satisfaction and retention of 1st-year teachers: The importance of effective school management. *Educational Administration Quarterly, 40*(5), 742–771.

Symeonidis, V. (2015). *The status of teachers and the teaching profession: A study of education unions' perspectives*. https://download.ei-ie.org/Docs/WebDepot/The%20Status%20of%20Teachers%20and%20the%20Teaching%20Profession.pdf

Tiplic, D., Brandmo, C., & Elstad, E. (2015). Antecedents of Norwegian beginning teachers' turnover intentions. *Cambridge Journal of Education, 45*(4), 451–474.

UNESCO. (2006). *Teachers and educational quality: Monitoring global needs for 2015*. http://unesdoc.unesco.org/images/0014/001457/145754e.pdf

Upton, M. G., & Egan, T. M. (2010). Three approaches to multilevel theory building. *Human Resource Development Review, 9*(4), 333–356.

Vandenberghe, C., & Tremblay, M. (2008). The role of pay satisfaction and organizational commitment in turnover intentions: A two-sample study. *Journal of Business and Psychology, 22*(3), 275–286.

Varga, M. (2017). The effects of teacher-student relationships on the academic engagement of students. *International Journal of Educational Research, 53*, 330–340.

Whipp, J. L., & Geronime, L. (2017). Experiences that predict early career teacher commitment to and retention in high-poverty urban schools. *Urban Education, 52*(7), 799–828.

Yang, G., Badri, M., Al Rashedi, A., & Almazroui, K. (2018). The role of reading motivation, self-efficacy, and home influence in students' literacy achievement: A preliminary examination of fourth graders in Abu Dhabi. *Large-Scale Assessments in Education, 6*(1), 1–19.

Dr. Lixia Qin is an Educational Policy researcher for the University of Wisconsin System and has more than 15 years of research experience in K-12 and higher education settings. She received her Ph.D. degree in Educational Administration at Texas A&M University (College Station) and has over 30 publications in the United States and China in a range of peer-reviewed journal articles, book chapters, and a book. Her research centers on the complex effects of school policy and organizational structures on teachers and students, particularly efforts to increase equity in student access and outcomes.

Chapter 17
Daycare Centers' Composition and Non-native Children's Language Skills at School Entry: Exploring the Nature of Context Effects Using Multilevel Modeling

Nina Hogrebe and Anna Marina Schmidt

Abstract Due to segregation in early childhood education, the demographic makeup of daycare centers varies considerably. At the same time, the daycare centers' composition (e.g., the proportion of children living in poverty) affects children's competences. However, as the exact relationship of this association is still unclear, we study the effect of preschool composition on children's language skills at school entry by exploring linear and nonlinear relationships. Using the example of a medium-sized German city, we combine data from a school entry examination and a preschool survey (7,604 children in 84 daycare centers) and employ multivariate latent and logistic regression analyses within a multilevel modeling framework. We find indications for nonlinear relationship between daycare centers' demographic composition and non-native children's grammar and German skills. More precisely, a negative effect of increasing risk compilation proportions is weakened when proportions of 30–40% are reached. Possible explanations for this turning point are discussed.

Keywords Segregation · Composition effects · Language skills · Non-native children

N. Hogrebe (✉)
Hamburg University of Applied Sciences, Hamburg, Germany
e-mail: Nina.Hogrebe@haw-hamburg.de

A. M. Schmidt
University of Münster, Münster, Germany

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_17

17.1 Introduction

Early Childhood Education and Care (ECEC)[1] is believed to contribute to educational equality by being especially effective for disadvantaged children (European Commission, 2011). As to this perspective, Vandenbroeck (2015) specifies that the children's competence development is linked to the demographic make-up of ECEC settings. According to the author, most positive effects are realized in mixed groups whereas a concentration of children from underprivileged or disadvantaged families disparages their learning outcomes.

Following this line of argumentation, it is problematic that the actual distribution of children in ECEC settings provides a stark contrast to this perspective. Segregation in ECEC is highly developed in many countries, and the demographic makeup of daycare centers varies considerably. Research for the USA and Germany, for example, shows that the proportion of poor and/or migrant children in ECEC settings varies between zero and more than 70%. This variation is not only visible between states but also within cities and city districts and exceeds segregation in kindergarten and primary school (e.g., Becker & Schober, 2017; Hogrebe, 2014, 2016; Potter, 2016).

Consequently, the demographic composition of daycare centers is often rather homogeneous in terms of children's linguistic, ethnic, and/or social background. According to a study from the USA, for instance, about 30% of preschoolers enrolled in public schools attend racially isolated schools, i.e., schools that have 90% or more students of color (Frankenberg, 2016). Disadvantaged children are particularly affected by this: More than half of all black and Hispanic preschool students as well as children lacking English proficiency (LEP) are enrolled in such racially isolated non-white schools. Additionally, non-white sub-groups are also generally quite segregated from each other (Piazza & Frankenberg, 2019).

This impression is supported by another study that is not only looking at public school-based programs but includes data on other center-based as well as home-based programs (Urban Institute, 2019). Basically, the study shows a u-shaped distribution of Black or Hispanic enrollment shares with only a few programs (less than 20%) having moderate black or Hispanic enrolment shares (i.e., 30–70%). Reid et al. (2015) describe a similar pattern in relation to poverty. They find that almost half of the children in their study sample (47.1%) attend high-minority classrooms (i.e., classrooms with 70–100% minority children) in which about 75% of the children are poor as well. Only very few children (17%) visit classrooms that are racially mixed and medium–high income. To the authors, a mixed demographic make-up in ECEC settings is "more the exception than the rule" (p. 7).

[1] ECEC encompasses different forms of regulated arrangements that provide education and care for children from birth to compulsory primary school age. In this paper, we focus on center-based out-of-home care of children aged one to six. The respective sites where such education and care offers are provided are labelled daycare centers or ECEC settings throughout the paper. The term preschool is used in the context of research from the USA and usually refers to children aged two to five.

Segregation patterns in other countries are different from those in the USA. In Germany, for example, there are more settings with relatively low proportions of migrant children and only few settings with rather high proportions. Still, daycare centers that are visited by migrant children are characterized by average proportions of children with a language other than German which are nearly two times higher as they are for non-migrant children (Hogrebe et al., 2021). This indicates that these children on average visit daycare centers with much higher concentrations of children who do not speak German at home, which–assumedly–puts them at a double risk.

17.2 Purpose of the Present Study

Although there is a growing evidence-base that demonstrates the influence of ECEC settings' composition on children's competence development, the specific nature of the relationship is not quite clear. So far, mostly the effect of a continuous increase in the proportion of disadvantaged children has been analyzed, but Becker and Schober (2017) argue that "it seems unlikely that the association between the share of children with certain background characteristics and children's cognitive and language development is linear" (p. 10). Therefore, the purpose of our study is to estimate the effect of the demographic composition of center-based ECEC on non-native children's language skills at school entry and, more precisely, to explore the specific nature of this association. Before we present our study results, we shortly depict our theoretical assumptions on the importance of the peer-related makeup of group settings for children's language development and shortly summarize previous studies on such context effects.

17.3 Background

17.3.1 Theoretical Framework

Socioecological and sociocultural theories on human development (e.g., Bronfenbrenner, 1990; Rogoff, 1990; Vygotski, 1978) emphasize the importance of social interactions in a meaningful and stimulating environment for child development. Although adults are often considered the primary interaction partners, children spend a great amount of time with near-age mates. In ECEC settings, children have the possibility to get in contact with peers of the same, younger, or older age who differ in their abilities and competencies. Such individual differences appear to have significant implications for peer interactions, and the daily interactions and play activities offer various learning opportunities. The older the children are, the more likely do they play in larger groups and are, therefore, affected by a larger number of children with different abilities (Andresen, 2005; Becker & Schober, 2017; Coplan &

Arbeau, 2009; Yarrow, 1975). Peers are often considered to be a specific stimulus for each other due to their relationships at eye level (Youniss, 1994). Following the concept of the "Zone of Proximal Development", Vygotski (1978) also frames peers as important interaction partners with the addition that they have to be more capable.

Despite such longstanding insights on the relevance of peers for children's cognitive development and the fact that children spend a lot of time in ECEC settings with other children, the potential of child–child interactions for language development has often been neglected and underestimated (Branco, 2005). In that regard, children can also promote each other in their language use. Already at the age of five, children can have adequate conversations with their agemates (Howes et al., 1992). Research on the associations between multilingual children's peer interactions and language growth in the surrounding language points to the linguistic stimulation potential of peer interactions in early educational settings (Gámez et al., 2019; Palermo et al., 2014; Rydland et al., 2014). In her group socialization theory, Harris (2009) even claims that the peer environment can exert a greater influence on language development than parental input.

Consequently, the higher the respective peers' competences, the better a child's language skills should be. Segregation and preschool composition are relevant in this regard because it is assumed that if several children with the same cultural background are present, they usually find themselves together (Schneider-Andrich, 2021, p. 68). As language is often connected with cultural background, this might influence the quantity and quality of the national language input for non-native language learners.

17.3.2 Summary of Previous Research on Composition Effects

As linguistic competencies are related to various family background characteristics such as the migration status respectively the language spoken at home or the families' socioeconomic or parental education status, for example, one might also assume that the composition of such demographic aspects in ECEC settings is a relevant influencing factor for language development. Respective context studies point to the importance of the average cognitive or linguistic abilities of peers and the ethnic and/or social composition of ECEC settings for children's language skills (e.g., de Haan et al., 2013; Fram & Kim, 2012; Henry & Rickman, 2007; Mashburn et al., 2009; Reid & Ready, 2013; Schechter & Bye, 2007). In those studies, a higher proportion of disadvantaged children in daycare centers is usually associated with lower language skills. Some studies suggest that a better or worse peer-based linguistic environment is especially an advantage respectively disadvantage for children with lower language skills (e.g., Justice et al., 2011; Niklas & Tayler, 2018).

Most of the research on composition effects in ECEC assume linear associations, and little is known about possible nonlinear relationships. Only Reid and Ready (2013) report that "diversity improves children's expressive language learning when neither Whites nor minorities are the overwhelming majority" (p. 1098) for racial or ethnic composition. Research from the German school context indicates that a negative correlation between the proportion of children with a migration background and children's language skills is only found for schools with a proportion of 40% or more (Stanat, 2006). So far, no such turning points or thresholds are established in the literature for ECEC (Becker & Schober, 2017).

17.3.3 Research Question

Against the background of the lack of knowledge on possible nonlinear context effects, we explore the specific nature of such context effects in ECEC. Using data from one example municipality in Germany, we use multivariate latent and logistic regression analyses within a multilevel modeling framework to identify the effect of preschool composition on non-native children's language skills at school entry. Considering different possible relationships, we test for linearity as well as nonlinearity.

17.4 Multilevel Modeling as an Analytical Strategy for Context Effects

Before we elaborate on the setup of our study, we briefly outline some basic principles of multilevel modeling (MLM) as an analytic strategy. Children in settings tend to be similar to each other due to selection processes, for example. This violates the assumption of independence of the observations in standard statistical tests (Hox, 2010; Snijders & Bosker, 2012). In order to account for this selective clustering of children in daycare centers, MLM is applied as an analysis strategy indicating a relationship of a group membership and individual characteristics.

17.4.1 Five-Step Analytical Process

Following Hox (2010), MLM is done in five steps. In the first step, the variance fraction of the outcome variable is explained by the units' membership in a context. Here, the population estimator (also: *fixed effects* because the estimator remains the same for the respective context) contains only the intercept, which is assumed to be variable between context units. This empty model is described as the *random-intercept-only model* and decomposes the variation of the outcome variable into its variation within groups as well as between groups.

In the second step, the causal effects of the exogenous individual characteristics are estimated for the population, controlling for context membership. In this *random-intercept model*, the exogenous variables are added to the model at micro-level. The effects of the variables are estimated in such a way that they do not vary between the contexts (*fixed effects*). Enders and Tofighi (2007) recommend to center metric variables and present two forms for this in MLM: Centering variables around the group mean (*centering within cluster*) and centering around the overall mean (*centering at the grand mean*). While macro-level variables should be centered around the overall mean, micro-level variables can be centered around both the group and the overall mean. This should be done depending on the research question: Centering around the group mean is recommended when micro-level variables are of interest and for cross-level interactions; centering around the overall mean is recommended when interactions between macro-level variables are investigated.

In the third step, we test whether the effect of the individual variables varies systematically between the context units, i.e., whether context-specific interaction effects can be observed. This *random-intercept-random-slope model* tests for context dependence as well as potential explainability by the exogenous macro-level variables. The variance of the micro-level variables as well as their covariance between context units are estimated. Basically, a distinction is made between models in which the *slope* coefficient is constant or *fixed* and models in which it can vary and, like the intercept, is modeled as *random.*

A fourth step follows when systematic differences in the mean value of the outcome of interest become apparent. In this *intercept-as-outcome model*, the variation in the expected value of the outcome variable across contexts is explained by a context feature. This model assumes the independence of the effects of the exogenous context and individual variables on the one hand and the joint additive effect of these effects on the outcome variable on the other hand.

The fifth analysis step completes the analytical MLM process and is applied when the effect of the individual variables systematically differs across contextual units. The *random-coefficient model* assumes that the context variables (macro-level) moderate the effect of the individual variables (micro-level) in the context. This means that both the individual and the context characteristics do not have an independent effect on the outcome variable, but interact with each other, i.e., they interact across levels (*cross-level interaction*).

17.4.2 Model Fits and Indices

To identify whether the data recommend MLM, first, the measure of the *intraclass correlation coefficient* (ICC) is calculated. The ICC describes the proportion of variance explained by context membership. Its value should be significantly greater than 0.05, indicating that the variance of the outcome variable can be explained at least in part by context membership.

If the variables at both levels are added, the theoretic measures *Akaike's Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC) test the model fit and compare different models at the descriptive level. The values should be smaller in comparison to the random-intercept-only model. Also, the measure R^2 is output when exogenous variables are added to the model and shows how much of the variance can be explained at each level–the larger, the more variance in the outcome variable can be explained by the added variables.

17.4.3 Assumptions and Data Requirements

In order to be able to apply MLM, the data must meet certain requirements: First, it must have a clustered structure, which means that individual characteristics are embedded in a grouped context (here, the rule of thumb of at least ten cases within each unit applies; otherwise, Snijders and Bosker (2012) recommend performing a covariance analysis). Furthermore, the outcome and exogenous variables are assumed to be dependent. Finally, the residuals are uncorrelated within the respective levels and independent of the explanatory variables as well as normally distributed, and their variance is homoscedastic (i.e., the variance of the residual errors is constant and independent of the explanatory variables' values). Multilevel analysis further assumes that the collection of variables is free of measurement error, the outcome variable is at least interval-scaled, and the exogenous variables at the micro- and macro-levels are nominally scaled with k manifestations in the form of $k - 1$ dummy variables and zero–one scaling (or plus/minus one for centered effects) (Langer, 2010; Snijders & Bosker, 2012).

17.5 Applying MLM: Composition Effects and Non-native Children's Language Skills

17.5.1 Sample

Using the example of a medium-sized city in Germany (population of about 300,000), data are taken from the school entrance examination (SEE) conducted from 2010/2011 to 2015/2016 which is obligatory for all children ($n_{\text{children/total}} = 14{,}333$). Besides performing a social pediatric developmental screening including language skills, the SEE also gathers information on background characteristics such as family status, migration, or other developmental risk-factors. As it is known which preschool a child attends and since when, it is possible to aggregate the data on setting level ($n_{\text{centers}} = 172$). A survey generates additional information on the daycare centers' structural quality and their approach to language instruction. Realizing a response rate of about 50% results in an effective sample size of 7,604 children in 84 daycare centers. (A comparison of the total with the realized sample does not indicate any bias. The results are not displayed here but are available on request).

17.5.2 Variables

17.5.2.1 Outcome Variables

The outcome of interest is children's language competencies at school entry. In the SEE these are evaluated using four standardized screening tests that consist of six to ten exercises, respectively: *articulation*, *pluralization*, *prepositions*, and *phonetics*. Additionally, the physicians assess the child's *grammar use* (0 = deficient, 1 = borderline, 2 = normal) and overall *German skills* (0 = none, 1 = some, 2 = good) as displayed in the communicative situation of the examination. Table 17.1 shows descriptive statistics for the six outcome variables. The results of an explorative factor analysis indicate that the variables *preposition* (0.886), *pluralization* (0.841), and *grammar use* (0.816) load on a common factor ($\alpha = 0.826$). This corresponds to the internal data structure shown in a validation study of the screening instrument (Petermann et al., 2009). We labeled this factor *grammar skills* and treated it as a latent variable in the analyses. For reasons of simplicity regarding the interpretation of the results, metric outcome variables were z-standardized.

Table 17.1 Descriptive Statistics of the Outcome Variables

Variable	**Values**	**M**	**SD**	**Min**	**Max**
Articulation	0–10	9.4	1.1	0	10
Pluralization	0–7	6.2	1.6	0	7
Preposition	0–8	6.6	1.8	0	8
Phonetics	0–6	5.3	1.1	0	6
		N	**Freq (%)**		
Grammar skills	0 = deficient	1,206	15.90		
	1 = borderline	1,430	18.80		
	2 = normal	4,968	65.30		
German skills	0 = none	175	2.30		
	1 = some	687	9.00		
	2 = good	6,742	88.70		

Note N = 7,604 children; M = Mean; SD = Standard Deviation; Min = Minimum; Max = Maximum

17.5.2.2 Covariates

On individual level (level 1), the SEE gathers different information on the child and its familial situation. We include several control variables related to language development: The child's *age* (in years) and *sex* (0 = female, 1 = male), *duration* (in years) and *weekly dosage* (25, 35, or 45 h) of preschool experience, and whether the child is *prematurely* born (0 = no, 1 = yes). In relation to family status we consider whether the child has an older *sibling* (0 = yes, 1 = no) and lives in a *single parent* household (0 = no, 1 = yes). The variable *mother tongue* reveals the child's first language (0 = German, 1 = [German and] other). Furthermore, the physician documents the *parents' German skills* as displayed in the communicative situation of the examination process (0 = some and good, 1 = none), the child's status regarding *preventive examinations* (0 = complete, [and only 1 missing] 1 = incomplete), and whether the child ever experienced any additional *non-formal educational activities* like swimming, music, or sports (0 = yes, 1 = no). Finally, the variable *cumulative risk* (0 = no, 1 = yes) is computed and relates to children to whom at least three of the above-mentioned risk indicators apply.

The preschool survey generated information on the level of daycare centers (level 2). We included the *staff-child-ratio* as an overall structural quality indicator and three variables as proxies for language-related process quality: (1) How often do *language-related topics* emerge *in team meetings* (less than every 3 months, one to two times in 3 months, once a month, more than once a month; each coded with 0 = no and 1 = yes), is *professional training on language instruction* provided (0 = no, 1 = yes), and do parent-teacher *co-operation* activities explicitly target language issues (0 = no, 1 = yes).

Table 17.2 shows descriptive statistics for all covariates. All continuous controls are grand-mean centered. If categorical variables are not dichotomous, dummy variables were computed.

17.5.2.3 Composition Variables

The SEE allows us to compute the demographic makeup of the daycare centers. Using the cohorts from 2010/2011 to 2015/2016, we retrospectively did so for the school years 2009/2010 to 2012/2013. On the basis of individual level variables, we computed four composition variables: the proportion of children (1) with migration background, (2) in need of language support, (3) without additional experience in non-formal education activities, (4) with incomplete preventive health examinations and (5) with premature birth. Additionally, the preschool survey provides information on (6) the proportion of children exempt from fees as an income-related indicator for poverty. Table 17.3 shows that the respective proportions in daycare centers at least vary from zero to 26.6% (incomplete preventive examinations) but may be as high as 92.0% (poverty).

However, correlation analyses point to the problem of multicollinearity between all composition variables (r = 0.789 to r = 0.966) except for premature birth (see

Table 17.2 Descriptive statistics of the covariates

Level 1	**Values**	**M**	**SD**	**Min**	**Max**	**Missing (%)**
Age	In years	5.72	0.39	3.25	7.58	0
Duration	In years	3.43	0.85	0.17	6.75	0
		N	**Freq (%)**			
Weekly dosage	1 = 25 h	448	5.90			0
	2 = 35 h	3,016	39.70			
	3 = 45 h	4,140	54.40			
Sex	0 = female	3,663	48.20			0
	1 = male	3,941	51.80			
Prematurely born	0 = no	6,581	85.70			0
	1 = yes	1,086	14.30			
Older Siblings	0 = yes	3,967	52.20			0
	1 = no	3,637	47.80			
Single parent household	0 = no	6,600	86.80			0
	1 = yes	1,004	13.20			
Mother tongue	0 = German	5,426	71.40			0
	1 = (German and) other	2,178	28.60			
Parents' German skills	0 = none	265	3.50			0
	1 = some	472	6.20			
	2 = good	6,867	90.30			
Preventive examinations	0 = complete	7,004	92.10			0
	1 = incomplete	600	7.90			
Non-formal educational activities	0 = yes	5,831	76.70			0
	1 = no	6,548	23.30			
Cumulative risk	0 = no	6,548	86.10			0
	1 = yes	1,056	13.90			
Level 2	**Values**	**Average**	**SD**	**Min**	**Max**	
Staff-child ratio		5.66	1.20	3.56	8.47	8.30
		N	**Freq (%)**			
Language-related topics team meetings						
less than every 3 months		5	6.00			2.40
one to two times in 3 months		27	32.10			
once a month		24	28.60			
more than once a month		26	31.00			

(continued)

Table 17.2 (continued)

Professional training	0 = no	16	19.00			1.20
	1 = yes	67	79.80			
Co-operation	0 = no	56	66.70			2.40
	1 = yes	26	31.00			

Note N = 7,604 children (level 1); N = 84 (level 2); M = Mean; SD = Standard Deviation; Min = Minimum; Max = Maximum; Freq. = Frequency

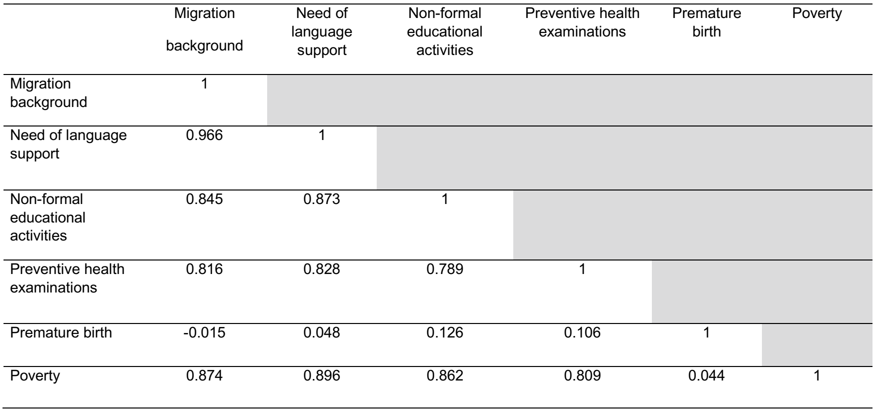
Table 17.3 Descriptive statistics of composition variables (in %)

	M	SD	Min	Max	Missing (%)
Migration background	27.01	20.53	0.00	81.08	0
Need of language support	20.40	17.13	0.00	66.54	0
Non-formal educational activities	22.83	20.96	0.00	77.74	0
Preventive health examinations	7.45	5.73	0.00	26.65	0
Premature birth	14.27	4.43	0.00	33.41	0
Poverty	24.69	28.46	0.00	92.00	0

Note N = 84 daycare centers; M = Mean; SD = Standard Deviation; Min = Minimum; Max = Maximum

Table 17.4). Based on the results of an explorative factor analysis (see Table 17.5), we decided to combine these five variables (Schneider, 2007). Consequently, we entered two predictors into our analyses: the proportion of premature birth as a manifest variable and a latently measured factor labeled "risk compilation" (M = 20.31, SD = 17.32, Min = 0.00, Max = 62.90).

Table 17.4 Test on multicollinearity



	Migration background	Need of language support	Non-formal educational activities	Preventive health examinations	Premature birth	Poverty
Migration background	1					
Need of language support	0.966	1				
Non-formal educational activities	0.845	0.873	1			
Preventive health examinations	0.816	0.828	0.789	1		
Premature birth	-0.015	0.048	0.126	0.106	1	
Poverty	0.874	0.896	0.862	0.809	0.044	1

Note Correlations between composition variables. Except for the correlation between premature birth and migration background, all correlations are significant ($p < 0.000$)

Table 17.5 Factor analysis of the composition variables

	Components	
	1	2
Need of language support	0.971	
Cumulative risk	0.971	
Migration background	0.966	
Non-formal educational activities	0.949	
Poverty	0.933	
Preventive health examinations	0.896	
Premature birth		0.999

Note Principal component analysis. Varimax with Kaiser-normalization. Cronbach's α with all components of extraction $1 = 0.949$

17.5.2.4 Missing Data

Missing values for individual level variables ranging from 0 to 6.4% were imputed using multiple imputation strategies with the multiple imputation module in SPSS 26. To account for missing data in the survey data, the full information maximum likelihood (FIML) estimator in MPlus 7 was used.

17.5.3 Modeling Approach

Two different models explore the specific relationship between preschool composition and children's language skills. Composition variables are entered as continuous variables in Model (1) to test for a linear relationship. In Model (2), a nonlinear relationship is tested to identify possible turning points at which the relationship changes. For this purpose, the squared proportions of the respective composition variables are added to the model. In both models, composition variables are divided by ten to indicate steps of 10%. A value of zero to five percent represents the reference group. All analyses are conducted in Mplus Version 7 (Muthén & Muthén, 1998–2015).

While the outcome variable *grammar skills* is entered as a continuous variable in a multivariate latent regression model, the categorical outcome variable *German skills* is employed in a logistic regression analysis. In contrast to linear regression models, in which the slope of the dependent variable is calculated by the increase of one unit of the predictor, probability ratios are calculated in logistic models. For this purpose, the coefficients are exponentiated by "e^{β}".

17.5.4 Results

In the *first step*, we examine which part of the variance of the outcome variables can be explained by the children's affiliation to a context, i.e., the corresponding ECEC center. The variables *articulation* and *phonetics* did not show sufficient variation on level 2 and were therefore excluded from the study. However, variance at the individual level is significant for the variables *grammar skills* and *German skills*, and variation due to daycare center affiliation also appears to explain differences in these language ability dimensions ($\beta = 0.246$ and $\beta = 1.179$ respectively, $p < 0.000$) (see Table 17.6). Also, the ICCs recommend the use of a multilevel analysis: For *grammar skills* 32.4% of the variance is explained by children's setting affiliation and for *German skills* it is 26.4%. For both outcome variables the observations are found to depend on their cluster membership, and differences in linguistic abilities are also due to differences between the clusters.

In the *second and third step* of the MLM, variables are added at the individual level. In step 2, the level 1 variables are modeled as fixed, i.e., each context (daycare center) has the same slope. By contrast, the modeling allows for different slopes between the clusters in step three. For *grammar skills*, coefficients in the two models hardly differ, but the standard deviations of two variables (children's first language and parents' German skills) double from the fixed to the random slope model indicating that the influence of the children's first language and their parents' German skills on children's grammar skills varies strongly between the facilities. Table 17.7 therefore

Table 17.6 Random-intercept-only model (empty model)

	Grammar skills		German skills		
	β	S.E	β	S.E	e^{β}
Grammar use	0.140*	0.051			
Preposition	0.111*	0.052			
Pluralization	0.095	0.049			
Threshold\$1[a]			−4.561***	0.167	0.010
Threshold\$2[a]			−2.732***	0.152	0.065
σ^2 e	0.513***	0.040	–	–	
σ^2 u_0	0.246***	0.055	1.178***	0.261	
ICC	0.324***	0.039	0.264***[b]	0.043	
AIC	51,728.178	98.503	5,633.621	28.110	
BIC	51,832.224	98.503	5,654.430	28.110	

Note Analytic sample (N = 7,604 children); number of clusters = 84; average cluster size = 90.52. β = beta coefficient, S.E. = standard error. e^{β} = odd. [a]In the logistic model, thresholds are output that differ only in sign from intercepts in linear regression models. The number of thresholds depends on the number of categories of the dependent variable (number of categories minus one). [b]In the logistic model, there is no variance at the within level. However, the ICC can be calculated using the formula σ^2 $u^0/(\pi^{2/3} + \sigma^2$ $u_0)$, where $\pi = 3.14159$

*$p < 0.050$, **$p < 0.010$, ***$p < 0.000$

Table 17.7 Random-intercept model

	Grammar skills		German skills		
	β	S.E	β	S.E	e^{β}
Age (in years)	0.112***	0.024	0.093	0.133	1.097
Duration (in years)	0.092***	0.015	0.546***	0.065	1.726
Sex	−0.013	0.016	−0.192	0.098	0.825
Prematurely born	−0.068*	0.028	0.158	0.191	1.171
Weekly dosage 35 h	0.007	0.031	0.091	0.176	1.095
Weekly dosage 45 h	0.047	0.031	0.255	0.188	1.290
Older siblings	0.033*	0.015	−0.199	0.104	0.820
Single parent household	−0.033	0.027	0.201	0.136	1.223
Mother tongue	−0.583***	0.037	−2.230***	0.142	0.108
Parent's German skills (none)	−0.787***	0.084	−2.475***	0.164	0.084
Preventive examinations	−0.540**	0.049	−0.714***	0.130	0.490
Non-formal educational activities	−0.941***	0.033	−0.793***	0.126	0.452
Cumulative risk	−0.253***	0.040	−0.452**	0.139	0.636
σ^2 e	0.292***	0.029	–	–	
σ^2 u_0	0.102**	0.038	0.296***	0.084	
ICC	0.259***	0.068	0.082***	0.022	
AIC	48,412.973	115.032	3,908.109	46.161	
BIC	48,634.939	115.032	4,019.092	46.161	
R^2 (within)	0.481[a]	–	0.472***	0.018	
R^2 (between)	–	–	–	–	

Note Analytic sample (N = 7,604 children); number of clusters = 84; average cluster size = 90.52. β = beta coefficient, S.E. = standard error. [a]Due to the modeling of the variables *first language* and *parent's German skills* as random, no model measures are output in MPlus except for AIC and BIC. R^2 is calculated using the following formula: 1 – (σ^2 e + σ^2 u_0 of the model)/(σ^2 e +σ^2 u_0 of the empty model) (see Snijders & Bosker, 2012, pp. 109–118)
*$p < 0.050$, **$p < 0.010$, ***$p < 0.000$

depicts the modified model for the outcome variable *grammar skills*, where the two mentioned individual level variables are modeled as random, while for the other level 1 variables the effect is assumed to be the same for all daycare centers. For logistic regressions, i.e., the outcome variable *German skills*, the regression constant is fixed, and coefficients always have the same values.

In terms of model fit, for both outcome variables the theoretic measures AIC and BIC are smaller compared to the empty models, indicating model modification. The R^2 describes that 48.1% of the variance can be explained by the level 1 variables for *grammar skills* and 47.2% for *German skills*, respectively. From a substantive perspective, the coefficients are consistent with expectations: Language-related characteristics such as the child's first language and the parents' German skills as well as education-related variables have the greatest influence.

For a possible explanation of the differences in slopes between clusters, we add in a *fourth step* predictors at the institutional level. Here, we calculate two different models: the composition variable is modeled as a linear relationship in Model (1) and a nonlinear relationship in Model (2).[2] In the *fifths and final step*, we additionally add interaction effects. For this purpose, the relationship between children's first language and their grammar skills is modeled as a random slope and regressed on the composition variable (*first language X composition*). According to Sommet and Morselli (2017), the calculation of interaction effects in a logistic regression is "a little more complicated" (p. 213). The authors propose a decomposition of the interaction: The effect of a level 2 predictor is estimated for each category of a level 1 variable in two dichotomously coded models. For our calculations, we estimate the effect of the composition variable on *German skills* of children with a first language other than (only) German in a first model and the effect for children with German as first language in a second model. For this purpose, we use two separate data bases and calculate the main effects with logistic regression. Here, results are presented for the subset of non-native children.

Tables 17.8 and 17.9 show the results of the fifth step of the multilevel analyses for *grammar skills* and *German skills*, respectively. With some exceptions, most of the level 2 co-variates do not significantly predict our outcome variables. This either means that the selected characteristics of the contextual framework are less relevant for the children's linguistic development, or the selected characteristics are insufficient proxies for mapping the quality of the daycare centers. With the exception of an interaction effect in the linear Model (1) for grammar skills, the same applies to the composition variable premature birth. For the sake of simplicity, we therefore focus on the latently measured factor risk compilation in the following.

In Model (1) the relationship of children's *grammar skills* and the centers' risk compilation appears to be a linear one, indicating that by an increase of 10% of children in risk compilation, all children's grammar skills are negatively affected by this ($\beta = -0.094$, $SE = 0.027$, $p < 0.010$) (see Table 17.8). However, we do not find a significant interaction effect in this model. By contrast, in Model (2) a nonlinear relationship emerges for children with not (only) German as first language. The results show that these children's *grammar skills* are affected nonlinearly by the risk compilation proportion in ECEC settings (linear coefficient: $\beta = -0.286$, $SE = 0.099$, $p < 0.010$; quadratic coefficient: $\beta = 0.041$, $SE = 0.014$, $p < 0.050$).

Focusing on *German skills* of children whose first language is not (only) German, no linear effect can be found in Model (1). In the same way, the model specification of the nonlinear relationship of setting composition and children's *German skills* in Model (2) identifies an increasing chance for children with not (only) German first language to get into the next higher category of German skills when they attend a center with a proportion of more than 40%. That is, if the risk compilation proportion increases by four units (each by 10% points), children have a slightly lower chance

[2] As we are especially interested in the effects for non-native children in this paper, which become visible when entering interaction effects in the next step, we do not present the results of this step here. However, they can be obtained from the authors upon request.

Table 17.8 Random-coefficient model (grammar skills)

	Model 1		*Model 2*	
	β	S.E	*β*	S.E
Individual characteristics				
Age (in years)	0.103***	0.023	0.104***	0.023
Duration (in years)	0.090***	0.015	0.092***	0.015
Sex	−0.014	0.016	−0.015	0.016
Prematurely born	−0.066*	0.028	−0.066*	0.028
Weekly dosage 35 h	0.007	0.030	0.005	0.030
Weekly dosage 45 h	0.047	0.030	0.047	0.030
Older siblings	0.033*	0.015	0.032*	0.015
Single parent household	−0.032	0.026	−0.034	0.026
Mother tongue	−0.253*	0.108	0.186	0.173
Parents ‘ German skills (none)	−0.897***	0.101	−0.896***	0.102
Preventive examinations	−0.210***	0,045	−0.210***	0.045
Non-formal educational activities	−0.428***	0.034	−0.427***	0.035
Cumulative risk	−0.246***	0.040	−0.245***	0.040
Institutional characteristics				
Staff-child-ratio	0.005	0.023	−0.011	0.024
Language-related topics in team meetings…				
…less than every 3 months	Reference category		Reference category	
…one to two times in 3 months	0.081	0.102	0.081	0.101
…once a month	0.077	0.107	0.079	0.108
…more than once a month	0.126	0.117	0.135	0.113
Professional training	−0.067	0.052	−0.075	0.052
Co-operation	−0.105*	0.051	−0.087	0.050
Composition variables				
Premature birth				
0–10%	Reference category		Reference category	
10% steps [Model (1) (2)]	0.019	0.060	0.197	0.168
quadratic [Model (2)]			−0.065	0.064
Risk compilation				
0–5%	Reference category		Reference category	
10% steps [Model (1) (2)]	−0.094**	0.027	0.056	0.076
quadratic [Model (2)]			−0.026	0.014
Cross-Level-Interaction				
Mother tongue × Premature birth				
0–10%	Reference category		Reference category	

(continued)

Table 17.8 (continued)

	Model 1		*Model 2*	
	β	S.E	*β*	S.E
10% steps [Model (1) (2)]	−0.159*	0.071	−0.548*	0.216
quadratic [Model (2)]			0.144*	0.063
Mother tongue × Risk compilation				
0–5%	Reference category		Reference category	
10% steps [Model (1) (2)]	−0.023	0.026	−0.286**	0.099
quadratic [Model (2)]			0.041*	0.015
σ^2 e	0.289***	0.028	0.289***	0.029
σ^2 u_0	0.041**	0.014	0.034**	0.012
σ^2 u_1 (interaction)	0.037**	0.011	0.028**	0.009
ICC	0.125***	0.032	0.106***	0.030
AIC	49,625.586	113.941	50,585.496	113.778
BIC	50,027.899	113.941	51,043.778	113.778
R^2 (within)[a]	0.565		0.574	
R^2 (between)[a]	0.833		0.862	

Note Analytic sample (N = 7,604 children); number of clusters = 84; average cluster size = 90.52. β = beta coefficient, S.E. = standard error. Latently modeled outcome variable grammar skills is z-standardized. Metric co-variates are grand-mean centered. Composition variables are entered as continuous variables in Model (1) and as quadric terms in Model (2). [a]Due to the modeling of the variables *first language* and *parent's German skills* as random, no model measures are output in MPlus except for AIC and BIC. R^2 (within) is calculated using the following formula: $1 - (\sigma^2$ e $+\sigma^2$ u_0 of the model)/(σ^2 e $+\sigma^2$ u_0 of the empty model). R^2 (between) is calculated with $1 - (\sigma^2$ u_0 of the model/σ^2 u_0 of the empty model) (see Snijders & Bosker, 2012, pp. 109–118)
$^*p < 0.050$, $^{**}p < 0.010$, $^{***}p < 0.000$

of having better German skills by a factor of 0.983 ($e^{-0.561+2*0.068*4}$) to 1. But their chances are better in settings with higher proportions: If the share is about 50%, the factor is 1.126 ($e^{-0.561+2*0.068*5}$) to 1. If the proportion increases to 70%, the factor is 1.478 ($e^{-0.561+2*0.068*7}$) to 1.

Thus, if non-native children attend preschools with particularly low or high proportions of children with socially disadvantageous characteristics, they show better *grammar skills* than corresponding children in preschools with medium proportions. In the same way, these children have nearly a one and a half times higher chance of having good or moderate *German skills* in contrast to no German skills at all in the latter daycare centers. Thus, from a certain point on, high proportions of children at risk in the settings increase non-native children's chance of having better language skills. Figure 17.1 depicts these relationships graphically and reveals that for both language outcomes the negative impact of higher risk compilation values turns when a proportion of about 40% is reached. Though the negative effect is weekend at this turning point, it is not completely leveled out, i.e., that preschools with very

Table 17.9 Random-coefficient model (German skills of non-native children)

	Model 1			*Model 2*		
	β	S.E	e^{β}	β	S.E	e^{β}
Individual characteristics						
Age (in years)	0.010	0.141	1.010	0.001	0.144	1.001
Duration (in years)	0.584***	0.071	1.793	0.595***	0.073	1.813
Sex	−0.299**	0.103	0.742	−0.288**	0.104	0.750
Prematurely born	0.206	0.150	1.229	0.200	0.150	1.221
Weekly dosage 35 h	0.130	0.190	1.139	0.146	0.191	1.157
Weekly dosage 45 h	0.308	0.187	1.361	0.318	0.187	1.374
Older siblings	−0.273*	0.107	0.761	−0.270*	0.107	0.763
Single parent household	0.178	0.153	1.195	0.200	0.154	1.221
Mother tongue	−2.188***	0.162	0.112	−2.181***	0.162	0.113
Parents' German skills (none)	−0.708***	0.132	0.493	−0.710***	0.133	0.492
Preventive examinations	−0.804***	0.130	0.448	−0.793***	0.131	0.452
Non-formal educational activities	−0.321*	0.135	0.725	−0.331*	0.136	0.718
Institutional characteristics						
Staff-child-ratio	0.117	0.075	1.124	0.124	0.082	1.132
Language-related topics in team meetings…						
…less than every 3 months	Reference category			Reference category		
…one to two times in 3 months	0.163	0.328	1.177	0.284	0.307	1.328
…once a month	0.629	0.343	1.876	0.772*	0.359	2.164
…more than once a month	0.617	0.343	1.853	0.670	0.409	1.954
Professional training	−0.167	0.198	0.846	−0.321	0.260	0.725
Co-operation	−0.572**	0.178	0.564	−0.543*	0.217	0.581
Composition variables						
Premature birth						
0–10%	Reference category			Reference category		
10% steps [Model (1) (2)]	−0.058	0.164	0.944	0.937	0.640	2.552

(continued)

Table 17.9 (continued)

	Model 1			*Model 2*		
	β	S.E	e^{β}	*β*	S.E	e^{β}
quadratic [Model (2)]				−0.276	0.196	0.759
Risk compilation						
0–5%	Reference category			Reference category		
10% steps [Model (1) (2)]	−0.089	0.050	0.915	−0.561*	0.221	0.571
quadratic [Model (2)]				0.068*	0.032	1.070
σ2 e	–	–		–	–	
σ2 u0	0.069	0.036		0.073	0.068	
AIC	3,975.840	0.001		4,881.561	3.008	
BIC	4,197.601	0.001		5,137.439	3.008	
R^2 (within)	0.345***	0.022		0.346***	0.022	
R^2 (between)	0.804***	0.133		0.968***	0.036	

Note N = 2,178 children in 78 clusters. *β* = beta coefficient, S.E. = standard error, e^{β} = odd. Metric co-variates are grand-mean centered. Composition variables are entered as continuous variables in Model (1) and as quadric terms in Model (2)
$^{*}p < 0.050$, $^{**}p < 0.010$, $^{***}p < 0.000$

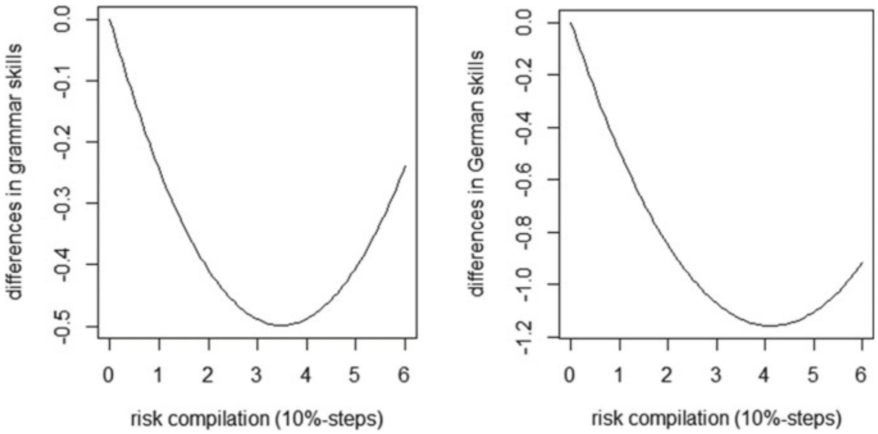


Fig. 17.1 Visual representation of the nonlinear correlation depending on children's first language (*Note* Left side: $N = 7{,}604$ children. The x-axis shows the proportion of children in risk compilation [in 10%-steps]. The y-axis shows the differences in children's grammar skills. Right Side: $N =$ 2,178 [only non-German speaking children]. The x-axis shows the proportion of children in risk compilation [in 10%-steps]. The y-axis shows the differences in children's German skills. Figure created with R)

low proportions of disadvantaged children still provide a more beneficial learning environment for non-native children in relation to their language skills.

17.6 Discussion

Our study generates findings on ECEC learning contexts and related structural inequalities and contributes to the understanding of the extent and educational relevance of segregation. By using an extensive set of composition variables and coming across the problem of collinearity, we showed that a disadvantaged demographic make-up of daycare centers is multidimensional. Although various studies on the relationship of daycare centers' composition and children's competences emphasize the importance of social and ethnic composition in daycare centers for children's skill development, they usually assume a linear relationship. In our study we also included a nonlinear relationship.

With regard to non-native children, which were at the focus of our analyses in this paper, such nonlinear effects of center composition and language skills become visible. According to our study results, a particularly low proportion of risk compilation seems to have the most favorable effect on both considered language dimensions. This finding is in line with previous research. However, by adding nonlinear associations to our analyses, we could additionally show that the negative effect of increasing risk compilation proportions is weakened when proportions of about 40% are reached.

A possible explanation for this turning point can be found in German ECEC policymaking that includes different mechanism to support daycare centers that operate in challenging areas or have an especially high intake of disadvantaged children. Most of these instruments direct more financial resources and/or personnel to such settings (see, for example, Anders et al., 2016). This might also explain the linear effect that was visible for all children's grammar skills as these children do not profit from such programs and resources.

Daycare center composition is the result of complex segregation processes. Therefore, it is unrealistic to create beneficial learning environments with very low proportions of disadvantaged children for all non-native children. Against this background, such needs-based and targeted programs and mechanisms seem to be a promising approach as they seem to be able to mediate some of the negative impacts of increasing risk compilations. That the negative effect is only reduced and not completely leveled out could point to direct and indirect peer effects operating. More research is needed on the mechanisms explaining such composition effects.

Acknowledgements This research was sponsored by the German Research Association (DFG) under the funding title HO 5561/1-1 and has partly been published in Hogrebe and Pomykaj (2019) and Pomykaj (2020).

References

Anders, Y., Rossbach, H.-G., & Tietze, W. (2016). Methodological challenges of evaluating the effects of an early language education programme in Germany. *International Journal of Child Care and Education Policy, 10*(9), 1–18.

Andresen, H. (2005). Role play and language development in the preschool years. *Culture and Psychology, 11*(4), 387–414. https://doi.org/10.1177/1354067X05058577

Becker, B., & Schober, P. S. (2017). Not just any child care center? Social and ethnic disparities in the use of early education institutions with a beneficial learning environment. *Early Education and Development.* https://doi.org/10.1080/10409289.2017.1320900

Branco, A. U. (2005). Peer interactions, language development and metacommunication. *Culture and Psychology, 11*(4), 415–429. https://doi.org/10.1177/1354067X05058580

Bronfenbrenner, U. (1990). The ecology of cognitive development. *Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 10*(2), 101–114.

Coplan, R. J., & Arbeau, K. A. (2009). Peer interactions and play in early childhood. In K. H. Rubin, W. M. Bukowski, & B. Laursen (Eds.), *Handbook of peer interactions, relationships, and groups* (pp. 143–161). The Guilford Press.

De Haan, A., Elbers, E., Hoofs, H., & Leseman, P. (2013). Targeted versus mixed preschools and kindergartens: Effects of class composition and teacher-managed activities on disadvantaged children's emergent academic skills. *School Effectiveness and School Improvement, 24*(2), 177–194. https://doi.org/10.1080/09243453.2012.749792

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138.

European Commission. (2011). *Early childhood education and care: Providing all our children with the best start for the world of tomorrow*. European Commission.

Fram, M. S., & Kim, J. (2012). Segregated from the start: Peer context in center-based child care. *Children &amp; Schools, 34*(2), 71–82. https://doi.org/10.1093/cs/cds011

Frankenberg, E. (2016). *Segregation at an early age.* http://school-diversity.org/wp-content/uploads/2016/10/Segregation_At_An_Early_Age_Frankenberg_2016.pdf

Gámez, P. B., Griskell, H. L., Sobrevilla, Y. N., & Vazquez, M. (2019). Dual language and English-only learners' expressive and receptive language skills and exposure to peers' language. *Child Development, 90*, 471–479.

Harris, J. R. (2009). *The nurture assumption: Why children turn out the way they do.* Free Press.

Henry, G. T., & Rickman, D. K. (2007). Do peers influence children's skill development in preschool? *Economics of Education Review, 26*(1), 100–112. https://doi.org/10.1016/j.econedurev.2005.09.006

Hogrebe, N. (2014). Indicators for a needs-based resource allocation in early childhood education: Regional data as valid proxies for setting level needs? *Journal for Educational Research Online, 6*(2), 44–65.

Hogrebe, N. (2016). *Choice and equal access in early childhood education and care: The case of Germany.* Paper presented at the ICMEC International Seminar Series. London: University of London. https://doi.org/10.17879/72229531014

Hogrebe, N., & Pomykaj, A. (2019). The school entrance examination as a data source for context studies in early childhood education and care. *Die Deutsche Schule, Beiheft, 14*, 71–86. https://doi.org/10.31244/dds.bh.2019.14.05

Hogrebe, N., Pomykaj, A., & Schulder, S. (2021). Segregation in early childhood education and care in Germany: Insights on regional distribution patterns using national educational studies. *Diskurs Kindheits- und Jugendforschung, 16*(1), 36–56. https://doi.org/10.3224/diskurs.v16i1.04

Howes, C., Droege, K., & Phillipsen, L. (1992). Contribution of peers to socialization in early childhood. In M. Gettinger, S. N. Elliott, & T. R. Kratochwill (Eds.), *Preschool and early childhood treatment directions* (pp. 113–150). Lawrence Erlbaum Associations.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). Routledge.

Justice, L. M., Petscher, Y., Schatschneider, C., & Mashburn, A. J. (2011). Peer effects in preschool classrooms: Is children's language growth associated with their classmates' skills? *Child Development, 82*(6), 1768–1777. https://doi.org/10.1111/j.1467-8624.2011.01665.x

Langer, W. (2010). Mehrebenenanalysen mit Querschnittsdaten (Multilevel analyses with cross-sectional data). In C. Wolf & H. Best (Eds.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (pp. 741–774). VS Verlag für Sozialwissenschaften.

Mashburn, A. J., Justice, L. M., Downer, J. T., & Pianta, R. C. (2009). Peer effects on children's language achievement during pre-kindergarten. *Child Development, 80*(3), 686–702. https://doi.org/10.1111/j.1467-8624.2009.01291.x

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Niklas, F., & Tayler, C. (2018). Room quality and composition matters: Children's verbal and numeracy abilities in Australian early childhood settings. *Learning and Instruction, 54*, 114–124. https://doi.org/10.1016/j.learninstruc.2017.08.006

Palermo, F., Mikulski, A. M., Fabes, R. A., Hanish, L. D., Martin, C. L., & Stargel, L. E. (2014). English exposure in the home and classroom. Predictions to Spanish-speaking preschoolers' English vocabulary skills. *Applied Psycholinguistics, 35*, 1163–1187.

Petermann, F., Daseking, M., Oldenhage, M., & Simon, K. (2009). *Sozialpädiatrisches Entwicklungsscreening für Schuleingangsuntersuchungen (SOPESS). Theoretische und statistische Grundlagen zur Testkonstruktion, Normierung und Validierung* (Social pediatric developmental screening for school entry examinations. Theoretical and statistical basics on test construction, standardization, and validation). Landesinstitut für Gesundheit und Arbeit des Landes Nordrhein-Westfalen.

Piazza, P., & Frankenberg, E. (2019). *Segregation at an early age: 2019 update*. https://cecr.ed.psu.edu/sites/default/files/Segregation_At_An_Early_Age_Piazza_Frankenberg_2019.pdf, 15 July 2021.

Pomykaj, A. (2020). *Sekundäranalysen in der Kindheitsforschung: Eine Diskussion der Chancen und Grenzen von Daten der Schuleingangsuntersuchung für Kontextstudien* (Secondary analyses in childhood research: A discussion of the opportunities and limitations of school entrance examination data for contextual studies). Universitäts- und Landesbibliothek Münster.

Potter, H. (2016). *Diversity in New York City's universal pre-k classrooms*. https://tcf.org/content/report/diversity-new-york-citys-universal-pre-k-...2

Reid, J. L., Kagan, S. L., Hilton, M., & Potter, H. (2015). *A better start. Why classroom diversity matters in early education*. The Century Foundation and The Poverty & Race Research Action Council. https://eric.ed.gov/?id=ED571023

Reid, J. L., & Ready, D. D. (2013). High-quality preschool: The socioeconomic composition of preschool classrooms and children's learning. *Early Education and Development, 24*(8), 1082–1111. https://doi.org/10.1080/10409289.2012.757519

Rogoff, B. (1990). *Apprenticeship in thinking. Cognitive development in social context*. Harvard University.

Rydland, V., Grøver, V., & Lawrence, J. (2014). The second-language vocabulary trajectories of Turkish immigrant children in Norway from ages five to ten: The role of preschool talk exposure, maternal education, and co-ethnic concentration in the neighborhood. *Journal of Child Language, 41*, 352–381.

Schechter, C., & Bye, B. (2007). Preliminary evidence for the impact of mixed-income preschools on low-income children's language growth. *Early Childhood Research Quarterly, 22*(1), 137–146. https://doi.org/10.1016/j.ecresq.2006.11.005

Schneider, H. (2007). Nachweis und Behandlung von Multikollinearität (Identification and handling of multicollinearity). In S. Albers, D. Klapper, U. Konradt, A. Walter, & J. Wolf (Eds.), *Methodik der empirischen Forschung* (pp. 221–236). Gabler. https://doi.org/10.1007/978-3-8349-9121-8_13

Schneider-Andrich, P. (2021). Early peer relationships and groups. A review of the state of research. *Frühe Bildung, 10*(2), 65–72.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology, 30*(1), 229–230.

Stanat, P. (2006). Schulleistungen von Jugendlichen mit Migrationshintergrund: Die Rolle der Zusammensetzung der Schülerschaft (School performance of immigrant youth: The role of student body composition). In J. Baumert (Ed.), *Herkunftsbedingte Disparitäten im Bildungswesen: differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit: Vertiefende Analysen im Rahmen von PISA 2000* (pp. 189–219). VS Verlag für Sozialwissenschaften.

Urban Institute. (2019). *Segregated from the start: Comparing segregation in early childhood and K-12 education.* https://www.urban.org/features/segregated-start

Vandenbroeck, M. (2015). Ethnic diversity and social inclusion in ECCE in Europe. In P. T. M. Marope & Y. Kaga (Eds.), *Investing against evidence. The global state of early childhood care and education* (pp. 105–118). UNESCO Publishing.

Vygotski, L. S. (1978). *Mind in society. The development of higher psychological processes.* Harvard University Press.

Yarrow, M. R. (1975). Some perspectives on research on peer relations. In M. Lewis & L. A. Rosenblum (Eds.), *Friendship and peer relations* (pp. 298–305). Wiley.

Youniss, M. (1994). *Soziale Konstruktion und psychische Entwicklung* (Social construction and psychological development). Suhrkamp.

Nina Hogrebe is a Professor of educational research with a special focus on childhood at the University of Applied Sciences Hamburg. Her main research areas are educational inequalities, early childhood education, care, educational systems, and governance. Currently, she runs a project funded by the Federal Ministry of Education and Research that investigates the role of daycare centers and their providers in early childhood education segregation. She is experienced with educational effectiveness studies and secondary analyses using both national and international comparative data.

Anna Marina Schmidt is a Senior Researcher at the University of Münster, Department of Education. She focused her Ph.D. on secondary analyses in childhood studies and is experienced in using different data sources (e.g., national and local data sets from the education and health sector). Her main areas of research are social inequalities in early childhood and secondary analyses in quantitative childhood research.

Chapter 18
Gender Effect at the Beginning of Higher Education Careers in STEM Studies: Does Female Recover Better Than Male?

Antonella D’Agostino, Giulio Ghellini, and Gabriele Lombardi

Abstract We explore if gender matter in the effect of the so-called “transfer shock” that in the literature is defined as a temporary decrease in academic performance by transfer students immediately following the transition to a new institution and the corresponding recovery prevalent for most students in succeeding semesters. Despite the fact that the gender issue is very relevant in Science, Technology, Engineering, and Mathematics (STEM) higher education, no gender analysis has been conducted in this particular framework. Nonetheless, as Italy experiments relevant migration flows of students between Secondary Education (SE) graduation and Higher Education (HE), we study the effect of transfer shock in this specific point of the students’ career from a gender perspective. Our econometric strategy refers to multilevel modelling that allows to take into consideration not only individual characteristics (i.e. gender) but also that STEM students are clustered into university courses. Using micro-data provided by the Italian Ministry of University and Research (MUR), we find that students moving from the southern to northern regions of the country for their higher education suffer a transfer shock, and gender matter in this specific context. Referring to our main results, we stress the importance of multilevel modelling in this framework.

Keywords STEM · Gender · Multilevel modelling · Academic performance · Higher education

18.1 Introduction

The importance of using multilevel modelling in educational studies has been widely addressed because it allows to appropriately model data that occur within multiple hierarchies, such as students within a certain classroom within a certain school or

A. D’Agostino (✉)
University of Napoli Parthenope, Napoli, Italy
e-mail: antonella.dagostino@uniparthenope.it

G. Ghellini · G. Lombardi
University of Siena, Siena, Italy

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_18

graduates nested in degree programmes nested in universities (Bock, 2014; Grilli & Rampichini, 2009; O'Connell & McCoach, 2008). The underlying idea is that each level should be a potential source of unexplained variability (Snijders & Bosker, 2012).

In this chapter, we show how multilevel modelling is a very useful tool in order to analyse the relationship between the internal student mobility and the student performance in STEM higher education studies (HE) in Italy from a gender perspective. This relationship is an interesting issue to be investigated for different reasons. First of all because Italy experiments relevant migration flows of students that are almost unidirectional from southern to northern/central regions (Attanasio & Enea, 2019; D'Agostino et al., 2019a; Enea, 2018). From this perspective, students who stay at home experience less pressure during their studies; namely, they face fewer "settling costs" from the economic (e.g. no expenses for board and lodging), psychological (e.g. no need to familiarize themselves with a new place and a completely different lifestyle), and social (e.g. less necessity of finding new friends) points of view. Thus, we can argue that the HE performance of movers can be negatively influenced by these stressors and these students could be penalized, even if they likely represent the part of the student population with the greatest spirit of initiative and enterprise.

Secondly, starting from 2014, Italian universities receive economic incentives from the Ministry of Education for providing degrees within the prescribed time period (Viesti, 2018), therefore the prediction of performance of students is an essential and challenging issue for them.

Finally, despite the fact that the gender issue is very relevant in STEM studies because there is currently a low proportion of women studying and graduating in STEM subjects (see among the others Cheryan et al., 2017; Enea & Attanasio, 2020), gender differences in STEM student's careers in Italy are still very poor and no studies analyse the intersectional effect of mobility and gender.

In this framework, the purpose of this analysis is to measure whether internal mobility differentiated by gender has an effect on the first-year performance of students enrolled in STEM programmes at bachelor degree level. Specifically, we adapted the well-known concept of a "*transfer shock*" introduced by Hills (1965) to this specific context. In the education literature, a transfer shock refers to a temporary decrease in academic performance by transfer students immediately following the transition to a new institution of higher education and the corresponding recovery prevalent for most students in succeeding semesters. Hence, our main research question is whether the gender matter in the relationship between mobility and performance in STEM at HE level, taking into account the context effect.

The study is based on MOBYSU.IT dataset that has been provided by the Italian Ministry of University and Research (MUR) thanks to an agreement between the Ministry and some Italian universities.[1] The chapter will proceed as follows: the second section summarizes the empirical framework of the analysis, the third section

[1] Database MOBYSU.IT [Mobilitá degli Studi Universitari in Italia], protocollo di ricerca MUR – Università di Cagliari, Palermo, Siena, Torino, Sassari, Firenze, Cattolica and Napoli Federico II, Coordinatore Scientifico Massimo Attanasio (UNIPA), Fonte ANS-MUR.

presents the data, variables, and the modelling strategy used, the fourth discusses our main results, and then we conclude.

18.2 Empirical Framework

18.2.1 Internal Student Mobility in HE in Italy

It is not possible to highlight the characteristics of Italian student mobility without relying on the fact that migration flows are unidirectional (from South to North), with very little probability of returning and enriching the specialized workforces of the areas of origin. The reason is mainly related to the higher reputation of the Italian universities located in the central/northern regions, which are linked to the greater job opportunities that these institutions can offer after graduation. Indeed, for Italian universities, a great part of their attractiveness is based on the relationships that they are able to establish within their own territories and with the external demand for their graduates that they are able to provide, two factors already very scarce in the southern area of the country (Petrosino & Schingaro, 2016). In general, the consequences of such transfers are that these students face more “settling costs” from an economic (e.g. expenses for board and lodging), psychological (e.g. the need to familiarize oneself with a new place and a completely different lifestyle), and social (e.g. a greater necessity to find new friends) perspective. All these costs are likely to be more pronounced in South to North transfers because of a variety of factors and background differences between the northern/central and southern regions.

Figure 18.1 shows how the concentration of bigger universities is sensibly higher in northern macro-areas, and the mobility is much bigger from South and Islands rather than the opposite, which is almost irrelevant in the migration dynamics.

As pointed out by Cersosimo et al. (2016), the decision of remaining in the area of origin is surely related to the convenience of remaining within one’s comfort zone, but it is also determined by the possibility of finding an educational offer that properly

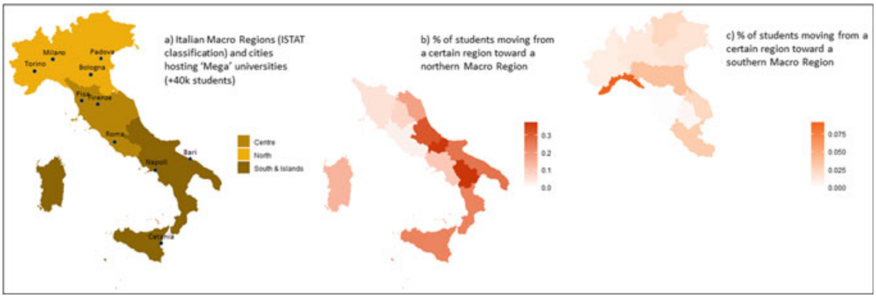

a) Italian Macro Regions (ISTAT classification) and cities hosting ‘Mega’ universities (+40k students)
Milano
Torino
Padova
Bologna
Pisa
Firenze
Roma
Bari
Napoli
Catania
Centre
North
South & Islands
b) % of students moving from a certain region toward a northern Macro Region
0.3
0.2
0.1
0.0
c) % of students moving from a certain region toward a southern Macro Region
0.075
0.050
0.025
0.000

Fig. 18.1 A synthetic representation of Italian geography and students’ movements for the a.y. 2014/2015 (*Source* Our elaboration on ANS-MUR data)

suits the student's needs. Apparently, only a small fraction of non-transfer students makes this decision because they feel they do not have any other choice.

Despite these indisputable challenges in deciding to move for higher education studies, some still hold the hypothesis that the choice of university can be driven by the desire to obtain a degree in the easiest way as possible. Actually, Argentin and Triventi (2015) find a significant difference between North and South in assessing marks at the Secondary Education level, with the latter being more "generous". From this point of view, the transition through a context less prone in providing "easy grades" can exacerbate the transfer shock. Nonetheless, De Paola (2008) and Lombardi and Ghellini (2019) investigate the (unsuccessful) attempts to attract students by softening grading policies made by those universities that are settled in the poorest regions and are consequently suffering from low demand by enrolling students, attributing much more importance to the conditions of students' destination territories. This idea is also supported by international studies on the internal mobility of students (Lörz et al., 2016). Indeed, several analyses show that the Italian student population is highly selective regarding the educational offers of the universities where they choose to apply (Cattaneo et al., 2018) and to the conditions of the job market they are trying to anticipate (Giambona et al., 2017). The combination of high-quality universities and healthy job markets—jointly able to guarantee the best chances of social mobility to future graduates—is the trump card for northern Italy (Bratti & Verzillo, 2019). At the same time, it is important to note once again how the presence of "better" universities in the northern regions is driven by a system of funding allocations that exacerbates the division between the two main macro-areas of the country and the consequent unidirectional migration flows (Cattaneo et al., 2017).

18.2.2 A Brief Overview on Students' Performance, Transfer Shock, and Gender

Different studies have specifically considered the performance of first-year students in STEM fields. The growing interest in these fields is causing a global increase in the adoption of strategies to stimulate STEM enrolment, which requires greater attention to the speed of students in adapting themselves to their new context. According to Lopez and Jones (2017), there is an evident stratification in who decides to enrol in STEM courses, since they find that the best explanatory variable for student performance in this area is the level of education of the father. Packard and Jeffers (2013) find it significantly necessary to accompany new STEM students in the transition to the higher education system to preserve their performance and motivation. Indeed, the lacking of self-esteem is one of the main factors damaging freshmen academic performance (especially concerning women), jointly with a learning environment too much oriented on the provision of concepts and instructions, rather than on the activation of stimulating ways of thinking (Schaeper, 2020).

On the other hand, the strong effect of social context in the STEM fields on student performance, regardless of the institution's facilities, is addressed in detail by Jackson (2010) and Jackson and Laanan (2015). In particular, difficulties in improving the performance of women in these fields are widely recognized and attributed to low socialization, a lack of mentors, and the internalization of stereotypical social norms. In the transition to a completely new institution, all these factors can cause—for both men and women—the abovementioned transfer shock. Keeley III and House (1993) analyse several factors that can lead to transfer shock. In particular, it seems that all transfer students generally experience some degree of transfer shock, especially men, who are on average outperformed by women. The youngest first-year students as well as the lowest performers at their previous institutions are usually the most affected by transfer shock. More specifically, Glass and Harrington (2002) find that transfer students seem to perform better in the long run than those who stay at their original university, but transfer students experience a drop in their performance during the first semester at their new university due to transfer shock. Also courses' characteristics can play a role in increasing performance. In particular, the most effective courses are those with a high percentage of women—well-recognized as better performing at university—and which require a greater number of work hours per week (Beekhoven et al., 2003; Bratti & Staffolani, 2013). Regarding STEM students specifically, Crisp et al. (2009) evidence how these fields are acquiring constantly greater importance in those public programmes stimulating higher education attendance, but it seems to present a difficulty in pursuing successfully a STEM degree by some categories of people, such as women or ethnic minorities. Also Soler et al. (2020) find lower performance for women in STEM area, even worsened at the Higher Education level with regard to the high school. From the transfer students' point of view, Cejda (1997) finds that first-year students in this area experience a stronger transfer shock than their colleagues in business, education, the fine arts and humanities, and the social sciences. This issue gains economic significance, especially in Italy, which is characterized by well-known unidirectional internal mobility of university students that is increasing the socio-economic gap between the northern and southern regions of the country (D'Agostino et al., 2019a). Even if studies about Secondary Education show how some family background predictors matter (Giambona and Porcu, 2015), studies at university level highlight that the performance of students does not seem to depend from socio-demographic features, but just on motivations and inclinations. Nonetheless, Non-resident students perform worse than residents, and male perform worse than females. High school final grade exhibits a positive effect on the performance both for good and very bad students, but Lyceum Southern students migrating in the North perform on average slightly better than Lyceum stayers (both southerners and northerners) (Adelfio & Boscaino, 2016; Adelfio et al., 2014; Boscaino et al., 2018; D'Agostino et al., 2021).

18.2.3 Using Multilevel Modelling in HE in Italy

In Italy, multilevel modelling has been especially used for assessing the relative effectiveness of educational institutions, namely the degree of achievements of their institutional targets (Grilli & Rampichini, 2009). From this point of view, the advantage of multilevel models lies in allowing students to be nested into institutions. In other words, outcomes defined at student-level can be used in order to measure the effectiveness of institutions themselves.

Another way to exploit courses’ quality through this class of models consists in exploring surveys and questionnaires for understanding students’ conditions and beliefs, distinguishing for the fact of attending different courses. As an example, Bassi et al. (2017) are able to draw useful insights about students’ satisfaction analysing their teachers’ evaluation. Interestingly, they are able to discover that course characteristics do not have strong importance, as didactic activities have. On the other side, Rampichini et al. (2004) and Bacci and Gnaldi (2015) use multilevel structure for accounting for the multidimensional nature of students’ satisfaction. Through this technique, they uncover the weaknesses of the courses. For example, courses which are bigger than the average size of their institution’s ones are less evaluated.

Nonetheless, an enormous potential emerges if students’ outcomes are analysed, far beyond the limited necessity of evaluating institutions. Thus, Meggiolaro et al. (2017) are able to study withdrawals, course changes, delays, and graduations of students using the course level in order to discover how fields of studies and size of each course are important for students’ university careers. Other studies employ multilevel modelling with regard to the path that students range across time. Indeed, when it is achievable to follow students also after their graduation it is also possible to evaluate their job market outcomes, important both for students and universities’ reputation. So, the success of students with similar academic backgrounds can be easily compared if their career is controlled also at the university and course level (Biggeri et al., 2001). On the other side, also the place in which the university is located is important. Even if there are marked differences between the effectiveness of universities and courses, the job market conditions of different geographical locations affect students’ chances of success (Bini et al., 2011). Following this idea, D’Agostino et al. (2019a, 2019b) analyse the determinants of student mobility by combining individual and contextual information through a multilevel approach, moving the perspective at the moment of choosing the university to be attended and highlighting the importance of Italy’s geographical characteristics. These results are confirmed also by the cross-country multilevel study by Agasisti and Cordero-Ferrera (2013), who compare Italy and Spain.

At this point, it is clear how multilevel modelling can express its potential also in analysing what happens at the very beginning of students’ enrolment. Indeed, Grilli et al. (2016) employ this class of models in order to evaluate the efficacy of pre-enrolment admission tests. Then, D’Agostino et al. (2020) interact internal mobility and career progression during the first two years, exploring the ability of multilevel models in analysing performance, as brought forward in the present chapter.

18.2.4 Measuring Academic Performance in This Study

As we stressed before, starting from 2014 Italian universities receive economic incentives from the Ministry of Education for providing degrees within the prescribed time period. The progress in academic curricula of university student is measured using the academic credits earned by students in each academic year. The credit system consists of the assignment of a certain credit when the student passes each course. He earns the credits which are based on that course. The students can earn credits according to his pace by taking any amount of time. The Laurea, which is equivalent to a Bachelor of Science in the European university system, is an undergraduate degree obtained after a three-year programme of study and it is strictly necessary to obtain 180 credits to accomplish it. One credit corresponds to a workload of about 25 h and the yearly workload for an average study course corresponds to about 60 credits which is equivalent to 1,500 h. In our data, for instance, the yearly average earned credits are about 43 ($SD = 14.62$) in the first academic year and about 33 credits ($SD = 16.26$) in the second year. From a policy perspective therefore it is more important to measure student academic performance in terms of the number of credits obtained by each student than on the score that they obtained for each exam. For this reason, the academic performance in this study is measured by the progress of the academic curricula and this measure cannot be considered a good proxy for the evaluation of each student, but it is a much better approximation for the number of exams successfully passed during the academic year. In other words, nothing can be said in our study about who the best students are in terms of score in this analysis but only in terms of progress in their academic curricula.

18.3 Data, Variables, and Method

18.3.1 Data

The database MOBYSU.IT covers the Italian university system as a whole. For this analysis, we used only the cohort of students enrolled in 2014/2015 in STEM programmes at bachelor's degree ($n = 51{,}821$). Official data from the Italian Ministry of Higher Education (MIUR) estimate that STEM students represent only 27% of the total number of first-year students, with a basically stable trend (approximately +0.3% annually over the last 10 years).

To classify degree courses into STEM fields, we used the definition provided by the EU Commission in 2015, based on Eurostat's Classification of Fields of Education and Training (Andersson and Olsson, 1999).

Accordingly, STEM fields of study can be classified into the following three macro-categories: (i) Natural sciences, mathematics, and statistics; (ii) Information and Communication Technologies; and (iii) Engineering, manufacturing, and construction.

Table 18.1 Student distribution by the three categories of STEM fields and gender (column %) in the academic year 2014/2015

Category	All sample	Male	Female
Natural sciences, mathematics, and statistics	37.17	28.37	51.37
Information and Communication Technologies	21.51	27.93	11.15
Engineering, manufacturing, and construction	41.33	43.70	37.49
Total	100	100	100
Number of students	38,773	23,938	14,835

Approximately, 13% of these first-year students drop out of the system, and approximately three per cent of them change course during the first academic year. Another 11% of these students had missing performance information. Therefore, the final database includes 38,773 first-year students enrolled in STEM fields in the academic year 2014/2015. According to our research objective, we selected only the first and second years of the students' careers since their enrolment.

Table 18.1 shows the percentage distribution of students by the three categories of STEM fields and gender. We observe some differences regarding the percentages of students in each category, such as a horizontal segregation that stresses how females are under-represented in two fields out of three.

The underlying structure of our database is complex because data contain longitudinal information grouped in several STEM degree courses. This leads to a three-level hierarchical data structure with 77,546 repeated measures (level 1) of 38,773 freshmen (level 2) nested in 658 university courses (level 3).

18.3.2 Variables

The response variable is the students' academic performance, which is measured using the academic credits earned in each academic year. In particular, for the purpose of the econometric analysis, we use a normal score transformation (Conover, 1999) of credit earned. The normal score transformation is designed to transform data so that it closely resembles a standard normal distribution. This transformation has been also used by Leckie (2013) for studying student performance in the UK. Let N be the total number of students in the dataset. First, the N observations are ranked based on their original scores. Then, the standard normal score CFU_j for the j-th ranked students is calculated as:

$$\text{CFU}_j = \Phi^{-1}\left[\frac{j-0.5}{N}\right], \tag{18.1}$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function. The advantage of this simple transformation is that it is order-preserving and

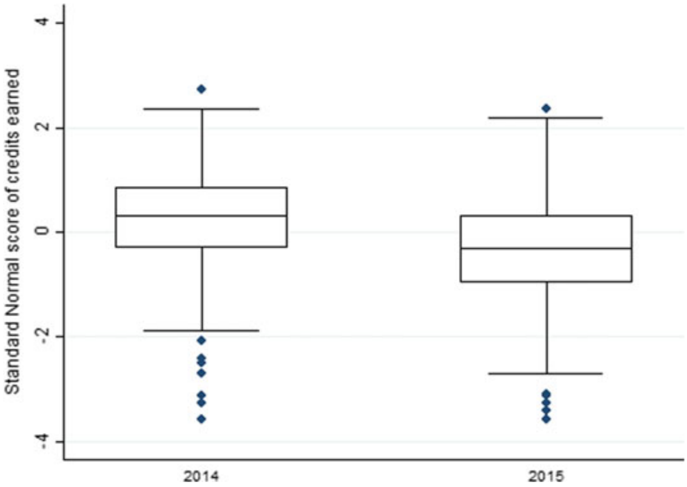


Fig. 18.2 Box plot of standard normal scores of credits earned by the year

students with the same number of credits will also receive the same standard normal score. Moreover, with this transformation, the effects of covariates can be interpreted in terms of standard deviations of the response variable. Last but not least, this transformation makes reasonable choice of econometric models based on a normal distribution assumption.

In Fig. 18.2, we reported the box plots of the standard normal scores of credits earned by year in order to make easier the comparisons of the two distributions. The lowest score of the normalized credit earned was about −2 in 2014, which is much higher than the lowest of scores in 2015. The average (median) of the scores in 2014 is higher than 0, whereas the average of scores in 2015 is lower than 0.

In the econometric model we control for several covariates, which definitions and summary statistics by gender are reported in Table 18.2. The main covariates of interest are the origin–destination movement of each student (henceforth called the mobility indicator) and gender. As the main object of this study concerns the effect of mobility from South to North, we use stayers and movers from the northern/central region as the baseline, comparing them with both movers and stayers from the South, separately.

Additionally, in our control strategy, the following variables are included. A dummy variable is calculated to identify if the final high school grade obtained was above the 75th percentile, which is another widely recognized predictor of student performance. Moreover, the type of high school attended is considered, using the scientific lyceum as a baseline and observing coefficients for students who attended a so-called classic lyceum or earned another diploma in secondary education. These categories are a well-known source of stratification at the higher education level (Ballarino & Panichella, 2016). Another binary indicator records if a student followed a "normal" road map, namely, if he/she enrolled in university later than the year after

Table 18.2 Descriptive statistics of individual explanatory variables

Variable name	Variable description	Female (%)	Male (%)	Total (%)
FEMALE	= 1 for female; = 0 for male	–	–	38.3
HSGRADE	= 1 if grade ≥ of 75th percentile of the HS grade distribution; = 0 otherwise	29.1	24.2	26.1
CLASSIC Lyceum (%)	= 1 if Classic Lyceum; = 0 otherwise	15.5	5.1	9.1
SCIENTIFIC Lyceum (%)	= 1 if Scientific Lyceum; = 0 otherwise	63.8	64.2	63.9
SCIENTIFIC Lyceum (%)	= 1 if other HS; = 0 otherwise	20.7	30.7	26.9
OTHER HS	= 1 if a student obtained his or her secondary education degree later than the year after the end of the high school; = 0 otherwise	5.00	4.80	4.9
STAY_SOUTH	= 1 stayers in southern regions; = 0 otherwise	31.5	25.5	27.8
MOV_SOUTH	= 1 movers from southern regions; = 0 otherwise	9.3	8.5	8.8
STAYMOV_N/C	= 1 stayers/movers in the north and central regions; = 0 otherwise	59.2	66.0	63.4
N		14,835	23,938	38,773

the end of high school. Indeed, as in other studies we expect that students who have entered higher education through a traditional track will progress more rapidly through their studies (Beekhoven et al., 2003).

Summary statistics highlight that freshmen in the sample are predominantly male (the proportion of females is only 38.3%), confirming the under-representation of women in STEM fields. Only 26.1% of them had an HS grade higher than 75th percentile value of the HS grade distribution. The majority came from scientific Lyceum. Only 4.9% enrolled in HE later than the year after the end of his/her HS. Finally, movers from south to north and central regions were only 8.8% of the total sample, whereas the stayers in South were 27.8%. Looking at the descriptive statistics by gender, some interesting consideration can be made. Women with a high HS grade were more than men (29.1% vs 24.2%). Approximately 15.5% of female students got a Classic Lyceum diploma, whereas that percentage for their male counterparts was only 5.1%. By contrast, male students that attained other Diplomas were more than females (30.7% vs. 20.7%).

From the percentages conducted by indicator mobility, there were also differences between genders. In particular, women in the group "Stayed and Moved North/Centre" are less than men, whereas they are more than men in the group "Stayed in South".

18.3.3 Method

The data used in this study have a natural hierarchical structure, namely, two repeated measures of credit earned by students that are nested in STEM courses (i.e. repeated measure, students, and courses are the level one, level two, and level three units of analysis, respectively). The repeated measures create time dependence between credit earned on two different occasions, and the students within the same university courses likely will have more characteristics in common than with students from other courses (i.e. measurements within-subjects and subjects within courses). Accordingly, in order to study the relationship between student mobility and performance, a multilevel linear regression model for continuous responses has been used as estimation strategy repeated measure, students, and courses are the level one, level two, and level three units of analysis, respectively. The computation of the intra-class correlation (ICC) allows us to measure the proportion of the total variance in the credit earned that is attributable to the two sources of variation (Goldstein, 2011; Snijders & Bosker, 2012).

The basic model (i.e. Model III (c) in Table 18.3) that tests the effect of a transfer shock in the overall population of freshmen and includes gender only as a main effect is specified as it follows:

$$\begin{aligned}\mathrm{CFU}_{tjk} =&\beta_0 + \beta_{\mathrm{TIME}}\mathrm{year}_{jk} + \beta_{SS}x_{jkSS} + \beta_{MS}x_{jkMS}\\ &+ \beta_{tSS}\mathrm{year}_{jk}\cdot x_{jkSS} + \beta_{tMS}\mathrm{year}_{jk}\cdot x_{jkMS}\\ &+ \sum_{h=1}^{H}\beta_h x_{hjk} + \sum_{l=1}^{3}\eta l AS_{kl} + v_k + u_{jk} + \varepsilon_{tjk}.\end{aligned} \qquad (18.2)$$

Equation (18.2) states, therefore, that CFU_{tjk} (normalized credit earned) in year t for student j (j = 1,…, J) in STEM course k (k = 1,…, K) is a linear function of student-level explanatory variables x, a time indicator variable (year—with the first as baseline), two dummy variables (AS) indicating different areas of study (Natural sciences, mathematics, and statistics, as the baseline) and a series of interaction effects. Finally, the error components ε_{tjk}, u_{jk}, and v_k are assumed to be mutually uncorrelated, i.i.d., and normally distributed with mean 0 and variance σ_ε^2, σ_u^2, and σ_v^2, respectively. It is important to stress that we estimate person-level effects because level 2 covariates are group-centred predictors.

In this setup, our main parameters of interest are β_{tMS} and β_{MS}. A statistically significant and negative estimate of β_{MS} indicates that movers from the South experience a transfer shock in the transition from high school to HE. Whereas a statistically significant and positive estimate of β_{tMS} suggests that these students are able to overcome the transfer shock in the second year of their STEM studies. In order to test the effect of gender further interaction effects are introduced into the model. Therefore, the more complex model (i.e. Model III (d) in Table 18.3) that tests if gender matter on the transfer shock includes further five parameters that represent the effect of interaction between gender and the mobility indicator, between gender and the time

Table 18.3 Estimate results—dependent variable is the normalized credits earned

	Model I			Model II	Model III			
	(a)	(b)	(c)		(a)	(b)	(c)	(d)
Year (baseline: First year of enrolment)				−0.568***	−0.568***	−0.568***	−0.568***	−0.568***
Gender (baseline: male)					−0.0359***	−0.0359***	−0.0359***	−0.042***
HS Grade (baseline: less than 75 percentile)					0.565***	0.563***	0.565***	0.565***
Classic Lyceum					−0.0993***	−0.0993***	−0.0991***	−0.0989***
Other Diploma					−0.227***	−0.227***	−0.226***	−0.226***
Late (baseline: enrolled HE less than one year after HS)					−0.144***	−0.144***	−0.144***	−0.144***
StaySouth					−0.071***	−0.058	−0.058	−0.057
MovSouth					−0.154***	−0.185***	−0.185***	−0.181***
Year#StaySouth						−0.0265	−0.0265	−0.0114
Year#MovSouth						0.0636***	0.0636***	0.0606***
Gender#Year								−0.0127
Gender#StaySouth								−0.0117
Gender#MovSouth								−0.176***
Gender#Year#StaySouth								0.284
Gender#Year#MovSouth								0.127***
Constant	−0.0561***	−0.0007***	−0.0568***	0.228***	0.170***	0.170***	0.182***	0.182***
Type of STEM course	No	No	No	No	No	No	Yes	Yes
Level 3: var(course)	0.121***	–	0.125***	0.121***	0.127***	0.127***	0.125***	0.125***
Level 2: var(student)	0.242***	0.371***	–	0.242***	0.258***	0.258***	0.258***	0.257***
Level 1: var(Residual)	0.616***	0.617***	0.857***	0.617***	0.455***	0.455***	0.455***	0.455***

(continued)

Table 18.3 (continued)

	Model I			Model II	Model III			
	(a)	(b)	(c)		(a)	(b)	(c)	(d)
ICC course	0.124	–	0.127	0.124	0.151	0.151	0.149	0.149
ICC student	0.371	0.375	–	0.371	0.448	0.458	0.457	0.457
- LogL	103,306	106,589	104,904	97,427	95,027	95,021	95,017	95,005
AIC	206,621	213,184	209,815	194,865	190,079	190,071	190,067	190,052
STEM courses (*K*)	658	–	658	658	658	658	658	658
Students (*J*)	38,773	38,773	–	38,773	38,773	38,773	38,773	38,773
Total observations (*N*)	77,546	77,546	77,546	77,546	77,546	77,546	77,546	77,546

Standard errors in parentheses, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

indicator and between gender, the mobility indicator, and the time indicator variable, respectively. A statistically significant estimate of some of the parameters of these interaction effects support the hypothesis that gender matter on the transfer shock.

18.4 Results

Our main results are presented in Table 18.3. We fit several models and both the Likelihood ratio (LR) tests and the Akaike's (1973) information criterion (AIC) were used in order to compare them and for the selection of the final model (Whittaker & Furlow, 2009).

In particular, the analysis was developed starting with the simplest models—Model I (a), (b), and (c)—and systematically moving towards more complex models (Model II–Model III).

The model without covariates, Model I (a), decomposes the total variation in the response variable into separate level-specific variance components. This model was fitted in order to detect whether there were statistically significant differences between repeated measures clustered into students and courses and therefore a multilevel model was required. The components of the variance are highly significant. The intercept—which measures the overall mean of the average number of credits (adjusted by the average in each degree course) earned by students—is statistically significant at the one percent level and negative. To test the overall significance of such model, a likelihood ratio test comparing the null random effect model to a null single-level model was conducted. The likelihood ratio test statistic was calculated as the difference in the -2*log-likelihood values for the two models, which in this case was statistically significant (LR test vs. linear model: $\chi^2(2) = 12{,}454.63$, p-value < 0.001).

The LR test also showed that this three-level model is preferred over Model I (b) which is a simpler longitudinal–two-level model ($\chi_1^2 = 6, 565.28$, p-value $<$ 0.0001). Moreover, it is also preferred over Model I (c), which is a clustered–two-level model ($\chi_1^2 = 3, 195.62$, p-value $<$ 0.0001).[2] Thus, there was evidence of longitudinal and course effects on credit earned, suggesting that a three-level multilevel model should be applied to take into account these differences.

The intra-class (courses) correlation equals 0.124 and intra-student correlation equals 0.371. Thus, approximately 12.4% of the variance is attributable to courses and 37.1% is attributable to students. The next step is the inclusion of the time indicator variable that is significant and negative (Model II), meaning that the number of credits earned in the second-year tends to decrease. The LR test ($\chi_1^2 = 11{,}758.48$, p-value <

[2] Snijders and Bosker (2012) discussed the technical issue that arises when using these LR tests in this framework because the null joint hypothesis is on the boundary of the parameter space. Indeed, we reject the null hypothesis even more strongly than we initially thought. Nevertheless, it is of more concern that when we do not just reject the null hypothesis based on the reported level, as in this situation, it is very likely that we should in fact reject the null hypothesis based on the actual level.

0.0001) and AIC confirm that the inclusion of this fixed effect improves the fit of the model. Model III (a) includes the main effect of student predictor variables described in Sect. 18.3. The LR test ($\chi_7^2 = 4{,}800.15$, p-value < 0.0001) and AIC confirm that the additional individual predictors improve the fit of the model. Then, Model III (b) adds the interaction term between the mobility indicator and the time indicator, and Model (c) includes a Level 3 covariate, namely the type of STEM course (the baseline is Natural sciences, mathematics, and statistics). For both models, the LR tests ($\chi_2^2 = 11.77$, p-value $= 0.0028$) and AIC show that Model III (b) provides a significant improvement in fit relative to Model III (a) and Model III (c) relative to Model III (b) ($\chi_2^2 = 7.38$, p-value $= 0.0250$).

This last model measures the effect of a transfer shock in the overall population of freshmen and includes gender as a main effect (see Eq. 18.2). Most of the results are as expected and consistent with the main findings in the literature. Turning to our principal research interest, the significant and negative sign of the coefficient β_{MS} indicates that southern students suffer from a transfer shock in the transition process from school to HE and the positive estimate of the coefficient of the interaction effect (β_{tMS}) clearly shows that movers from the South recover part of the credits they lost during the transition they experienced as first-year students.

Looking at the effect of gender, further interesting conclusions can be made. We reject the null hypothesis that the regression coefficient vector of interest is the same for both genders, because of the statistical significance of all parameters of interaction terms (see Model III (d)). In addition, both LR test ($\chi_5^2 = 25.73$, p-value < 0.0001) and AIC provide a significant improvement in fit relative to the Model III (c).

Findings suggest that movers, either males or females, appear to perform worse than their northern colleagues in the first year (the estimated coefficient of the variable *MovSouth* is and −0.185) and this effect is significant only for women (estimated coefficient of the interaction effect *Gender#MovSouth* is −0.176). In addition, the coefficient of the interaction effect "*Year#MovSouth*" is significantly different from zero and positive, the coefficient of the interaction effect "*Gender#Year#MovSouth*" is also significantly different from zero and positive, therefore we can argue that women recover their disadvantage in credit earned with respect to stayers/movers of North/Centre during their first-year of HE career better than men.

Finally, the effect of control variables is as we expected. The top-performing students come from scientific lyceums. Beyond the role of the type of HS as control for social stratification as explained in Sect. 18.3, this results confirms that scientific lyceum is a good training for STEM studies in HE. In addition, those who have a good previous school career and apply on time for a STEM degree course are more likely to outperform their peers during their university career.

18.5 Conclusion

The principal aim of this study was to explore the relationship between STEM students' mobility from the southern to northern regions of Italy and their academic performance at university. Not only we attempted to discover whether these students are affected by a transfer shock but we explored if this transfer shock also varies by gender. This was done by considering students' first- and second-year performance as measured by credits earned. Multilevel modelling was used in order to take into account the hierarchical structure of the data.

Empirical evidence showed that STEM students moving from southern regions to northern universities experience a transfer shock between their first and second academic year, confirming the results suggested by several papers on this issue (i.e. transfer students recover from transfer shock within a year) as it was stressed in Sect. 18.2. Moreover, we also found that gender matters in this particular transition because women seem to recover more than men in their academic performances in the second year whether they are movers or stayers.

These results are particularly meaningful and they might suggest the implementation of specific measures to mitigate the issue of transfer shock, especially for men. Obviously, further studies are needed to examine in detail what particular stressors transfer students face and which resources universities can provide for helping transfer students. Nevertheless, some interesting considerations can be made.

Referring to the results of Jackson and Laanan (2015), for instance, transfer students are more likely to successfully adjust to their university if they view their professors as approachable, accessible, and interested in their academic development. Therefore, we can argue that northern universities can move in this direction. Furthermore, universities can establish, for instance, learning communities that require the participation of all transfer first-year students from southern regions. Here, stronger students with solid study skills can assist their peers, share their best practices, and perhaps form study groups.

In addition, it is interesting to note that some studies conducted in the U.S. higher education context concluded that living on campus tends to improve student performance (De Araujo & Murray, 2010; Pascarella et al., 1993). Therefore, Italian universities can take note and improve their residential housing policies by creating and/or increasing student residential housing that is comfortable and thus providing opportunities for transfer students to have better academic performance than their counterparts residing in private houses.

From our findings, we can also argue that the needs of transfer students may differ by gender, since female transfer students show better resilience after one year. Unfortunately, we do not have information to investigate the reasons that can explain this gap. However, some further considerations can be made from a gender perspective. Even if women are under-represented in STEM universities courses, as stressed by several empirical studies (see among the other, Blackburn, 2017), they seem to be faster in recovering from the transfer shock they suffer, making a new contribution to the stream of literature finding that women potentially can outperform males also in STEM studies. This conclusion should contribute to convince both the policymaker and the academic community about the importance of moving towards STEM courses less damaged by the well-recognized gender inequality they are suffering nowadays.

This study has limitations that moderate the discussion of its findings and its implications for practice. First, this study used a single cohort of STEM first-year students. However, it is likely that the findings can be generally confirmed because different cohorts operate in similar settings in the Italian context. Thus, a replication of this study that focuses on more than one STEM cohort would shed light on our understanding of this issue. A second limitation is the simple data transformation used (see Eq. 18.2) in the econometric model. Since a more complex approach for such data exists in the literature (see, for instance, Grilli et al., 2016), using such an approach as a robustness check could be an interesting research topic for the near future.

References

Adelfio, G., & Boscaino, G. (2016). Degree course change and student performance: A mixed-effect model approach. *Journal of Applied Statistics, 43*(1), 3–15.

Adelfio, G., Boscaino, G., & Capursi, V. (2014). A new indicator for higher education student performance. *Higher Education, 68*(5), 653–668.

Agasisti, T., & Cordero-Ferrera, J. M. (2013). Educational disparities across regions: A multilevel analysis for Italy and Spain. *Journal of Policy Modeling, 35*(6), 1079–1102.

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* In B. N. Petrov & F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory (pp. 267–281). Akademiai Kiado.

Andersson, R., & Olsson, A. K. (1999). *Fields of education and training manual.* Manual for ISCED, 97.

Argentin, G., & Triventi, M. (2015). The North-South divide in school grading standards: New evidence from national assessments of the Italian student population. *Italian Journal of Sociology of Education, 7*(2), 157–185.

Attanasio, M., & Enea, M. (2019). *La mobilità degli studenti universitari nell'ultimo decennio in Italia* (pp. 43–58). Bologna, Il Mulino. ISBN: 978-88-15-28018-3.

Bacci, S., & Gnaldi, M. (2015). A classification of university courses based on students' satisfaction: An application of a two-level mixture item response model. *Quality & Quantity, 49*(3), 927–940.

Ballarino, G., & Panichella, N. (2016). Social stratification, secondary school tracking and university enrolment in Italy. *Journal of the Academy of Social Sciences, 11*(2–3), Social Inequality.

Bassi, F., Grilli, L., Paccagnella, O., Rampichini, C., & Varriale, R. (2017, June). New insights on student evaluation of teaching in Italy. In *Convegno della Società Italiana di Statistica* (pp. 263–274). Springer.

Beekhoven, S., De Jong, U., & Van Hout, H. (2003). Different courses, different students, same results? An examination of differences in study progress of students in different courses. *Higher Education, 46*(1), 37–59.

Bini, M., Grilli, L., & Rampichini, C. (2011). Contextual factors of the external effectiveness of the university education: A multilevel approach. *Italian Journal of Applied Statistics, 23*(1), 51–65.

Biggeri, L., Bini, M., & Grilli, L. (2001). The transition from university to work: A multilevel approach to the analysis of the time to obtain the first job. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 164*(2), 293–305.

Blackburn, H. (2017). The status of women in STEM in higher education: A review of the literature 2007–2017. *Science & Technology Libraries, 36*(3), 235–273.

Bock, R. D. (Ed.). (2014). *Multilevel analysis of educational data.* Elsevier.

Boscaino, G., Adelfio, G., & Sottile, G. (2018). *A distribution curves comparison approach to analyze the university moving students performance*. SIS2018: 49th Scientific Meeting of the Italian Statistical Society.

Bratti, M., & Staffolani, S. (2013). *Student time allocation and educational production functions*. Annals of Economics and Statistics/ANNALES D'E´CONOMIE ET DE STATISTIQUE, pp. 103–140.

Bratti, M., & Verzillo, S. (2019). The 'gravity' of quality: Research quality and the attractiveness of universities in Italy. *Regional Studies, 53*(10), 1385–1396.

Cattaneo, M., Horta, H., Malighetti, P., Meoli, M., & Paleari, S. (2018). The relationship between competition and programmatic diversification. *Studies in Higher Education, 44*(7), 1222–1240.

Cattaneo, M., Malighetti, P., Meoli, M., & Paleari, S. (2017). University spatial competition for students: The Italian case. *Regional Studies, 51*(5), 750–764.

Cejda, B. D. (1997). An examination of transfer shock in academic disciplines. *Community College Journal of Research and Practice, 21*(3), 279–288.

Cersosimo, D., Ferrara, A. R., & Nisticò, R. (2016). La mobilità geografica: da Sud a Nord senza ritorno. In G. Viesti (Ed.), *Università in Declino - Un'indagine degli Atenei da Nord a Sud*. Fondazione RES, Donzelli Editore.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin, 143*(1), 1.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). Wiley.

Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a stem degree: An analysis of students attending a Hispanic serving institution. *American Educational Research Journal, 46*(4), 924–942.

D'Agostino, A., Ghellini, G., & Lombardi, G. (2020). First and Second Year Careers of STEM Students in Italy: A Geographical Perspective. In *Book of short papers SIS 2020* (pp.1460–1465). Pearson.

D'Agostino, A., Ghellini, G., & Lombardi, G. (2021). Movers and stayers in STEM enrollment in Italy: Who performs better? *Genus, 77*(1). https://doi.org/10.1186/s41118-021-00141-7

D'Agostino, A., Ghellini, G., & Longobardi, S. (2019a). Out-migration of university enrolment: The mobility behaviour of Italian students. *International Journal of Manpower, 40*(1), 56–72.

D'Agostino, A., Ghellini, G., & Longobardi, S. (2019b). Exploring determinants and trend of STEM students internal mobility. Some evidence from Italy. *Electronic Journal of Applied Statistical Analysis (EJASA), 12*(4), 826–845.

de Araujo, P., & Murray, J. (2010). Estimating the effects of dormitory living on student performance. *Economics Bulletin, 30*(1), 866–878.

De Paola, M. (2008). *Are easy grading practices induced by low demand? Evidence from Italy* (MPRA Paper No. 14425). University Library of Munich, Germany.

Enea, M. (2018). From South to North? Mobility of Southern Italian students at the transition from the first to the second level university degree. In C. Perna, M. Pratesi, & A. Ruiz-Gazen (Eds), *Studies in theoretical and applied statistics*. Springer.

Enea, M., & Attanasio, M. (2020). Gender differences in Italian STEM degree courses: A discrete-time competing-risks model. In N. S. Alessio Pollice (a cura di), *Book of short papers—SIS 2020* (pp. 385–390). Pearson.

Giambona, F., & Porcu, M. (2015). Student background determinants of reading achievement in Italy. A quantile regression analysis. *International Journal of Educational Development, 44*, 95–107.

Giambona, F., Porcu, M., & Sulis, I. (2017). Students mobility: Assessing the determinants of attractiveness across competing territorial areas. *Social Indicator Research, 133*, 1105–1132.

Glass, J. J. C., & Harrington, A. R. (2002). Academic performance of community college transfer students and "native" students at a large state university. *Community College Journal of Research and Practice, 26*(5), 415–430.

Goldstein H. (2011). *Multilevel statistical models* (4th ed.). Wiley Series in Probability and Statistics.

Grilli, L., & Rampichini, C. (2009). Multilevel models for the evaluation of educational institutions: A review. In *Statistical methods for the evaluation of educational services and quality of products* (pp. 61–80). Physica.

Grilli, L., Rampichini, C., & Varriale R. (2016). Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: An approach based on quantile regression for counts. *Statistical Modelling, 16*(1), 47–66.

Hills, J. R. (1965). Transfer shock: The academic performance of the junior college transfer. *The Journal of Experimental Education, 33*(3), 201–215.

Jackson, D. L. (2010). *Transfer students in STEM majors: Gender differences in the socialization factors that influence academic and social adjustment* (Unpublished doctoral dissertation). Educational Leadership and Policy Studies, Iowa State University.

Jackson, D. L., & Laanan, F. S. (2015). Desiring to fit: Fostering the success of community college transfer students in STEM. *Community College Journal of Research and Practice, 39*(2), 132–149.

Keeley III, E. J., & House, J. D. (1993, May 16–19). *Transfer shock revisited: A longitudinal study of transfer academic performance.* Paper presented at the 33rd Annual Forum of the Association for Institutional Research, Chicago, IL.

Leckie, G. (2013). *Module 11: Three-level multilevel models—Concepts.* LEMMA VLE Module 11, 1–47. http://www.bristol.ac.uk/cmm/learning/course.html

Lombardi, G., & Ghellini, G. (2019). The effect of grading policies on Italian universities' attractiveness: A conditional multinomial logit approach. *Electronic Journal of Applied Statistical Analysis, 12*(04), 801–825.

Lörz, M., Netz, N., & Quast, H. (2016). Why do students from underprivileged families less often intend to study abroad? *Higher Education, 72*(2), 153–174.

Lopez, C., & Jones, S. J. (2017). Examination of factors that predict academic adjustment and success of community college transfer students in stem at 4-year institutions. *Community College Journal of Research and Practice, 41*(3), 168–182.

Meggiolaro, S., Giraldo, A., & Clerici, R. (2017). A multilevel competing risks model for analysis of university students' careers in Italy. *Studies in Higher Education, 42*(7), 1259–1274.

O'Connell, A. A., & McCoach, D. B. (Eds.). (2008*). Multilevel modeling of educational data.* IAP.

Packard, B.W.-L., & Jeffers, K. C. (2013). Advising and progress in the community college stem transfer pathway. *NACADA Journal, 33*(2), 65–76.

Pascarella, E., Bohr, L., Nora, A., Zusman, B., Inman, P., & Desler, M. (1993). Cognitive impacts of living on campus versus commuting to college. *Journal of College Student Development, 34*, 216–220.

Petrosino, D., & Schingaro, N. (2016). I cambiamenti dell'offerta formativa. In G. Viesti (Ed.), *Università in Declino - Un'indagine degli Atenei da Nord a Sud*. Fondazione RES, Donzelli Editore.

Rampichini, C., Grilli, L., & Petrucci, A. (2004). Analysis of university course evaluations: From descriptive measures to multilevel models. *Statistical Methods and Applications, 13*(3), 357–373.

Schaeper, H. (2020). The first year in higher education: The role of individual factors and the learning environment for academic integration. *Higher Education, 79*(1), 95–110.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modelling.* Sage.

Soler, S. C. G., Alvarado, L. K. A., & Nisperuza, G. L. B. (2020). Women in STEM: Does college boost their performance? *Higher Education, 79*(5), 849–866.

Viesti, G. (2018). *La laurea negata: le politiche contro l'istruzione universitaria.* Gius. Laterza & Figli Spa.

Whittaker, T. A., & Furlow, C. F. (2009). The comparison of model selection criteria when selecting among competing hierarchical linear models. *Journal of Modern Applied Statistical Methods, 8*(1), 15.

Antonella D’Agostino is Associate Professor of Economics Statistics at the Department of Management and Quantitative Studies, University of Napoli “Parthenope.” She holds a Ph.D. in Applied Statistics from Florence University. Currently, some of her research interests cover student mobility and performance issues. She participated in several international research projects. She worked as a consultant for Eurostat in statistical services in income, poverty, and social exclusion and for the International Labour Organization (ILO).

Giulio Ghellini is Full Professor in Social Statistics at the Department of Economics and Statistics, University of Siena. He has directed several projects for Eurostat related to survey design and data quality of E.U. Statistics on Income and Living Conditions and E.U. Social Survey. His current research interests are focused on Educational System Evaluation, Youth Transition to work, and Educational Mobility.

Gabriele Lombardi is Lecturer in “Social Research Methods” at the Department of Economics and Statistics of the University of Siena, where he held his Ph.D. in Economics specializing in Social Statistics of the student population at the Higher Education level. His main research areas are internal migration for Higher Education studies, geographical inequalities in Italy, and school-to-work transition.

Chapter 19
Service Satisfaction and Service Quality: A Longitudinal and Multilevel Study of User Satisfaction with Kindergartens in Norway

Håvard Thorsen Rydland, Åsta Dyrnes Nordø, and Dag Arne Christensen

Abstract Policymakers generally assume a simple and direct link between service quality and user satisfaction. However, research has focused on the absence of such a direct link—satisfaction is not a simple reflection of public authorities' service provision. We have surprisingly limited insight into this possible service quality paradox, and into which service characteristics that affect user satisfaction. This chapter studies if such a service quality paradox exists within Norwegian kindergartens. Kindergartens are considered a vital part of the Norwegian educational services. They have a legal obligation to both «safeguard the children's need for care and play» as well as «promote learning and formation as a basis for an all-round development». User satisfaction surveys have become more common within public administration, both as a quality measure and as a governance tool. This also includes educational services, with surveys directed toward both parents and pupils. The chapters test if quality measures (staffing and staff's education) impact satisfaction, and whether specific users (parents with the youngest children) are especially sensitive to changes in service quality. To do this we employ a novel dataset that combines individual-level data from four waves of the Norwegian kindergarten survey (2016–2019) and kindergarten-level panel data on service characteristics. The data enables us to estimate both within- and between-unit effects at the kindergarten level.

19.1 Introduction

What explains the demand and the support for welfare state services? Is there a one-to-one relationship between the quality of such services and user satisfaction?

H. T. Rydland (✉) · Å. D. Nordø · D. A. Christensen
NORCE Social Science - Norwegian Research Centre, Bergen, Norway
e-mail: hary@norceresearch.no

Å. D. Nordø
e-mail: asno@norceresearch.no

D. A. Christensen
e-mail: dach@norceresearch.no

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_19

Welfare services can be seen as resulting from a set of exchange relationships (Easton, 1965). Looking at it from the input side citizens—directly and through politicians and the media—make demands on the welfare system and provide support and resources in exchange. The welfare state meets such demands by providing policy outputs. This chapter focuses on the relationship between the output and input side of the welfare state. To do this we study two important feedback mechanism; objective outputs, consisting of hard indicators measuring service quality, and subjective inputs, consisting of soft indicators measuring user satisfaction (Bouckart & Van de Walle, 2003; Lindén et al., 2017). Even though user surveys have become standard tools for evaluating public services (Walle, 2018; Van Ryzin, 2011), research often reports odd results. The quality paradox, that is, the absence of a direct link between service quality and user's perception of such quality, is one example (Friman & Fellesson, 2009). We address this service quality paradox head-on by combining users' subjective evaluations of Norwegian Kindergartens with administrative data measuring service quality at the kindergarten level.

This chapter provides a thorough multilevel regression analysis to test the impact of kindergarten service characteristics on service satisfaction. We combine individual-level data from four waves (2016–2019) of the Norwegian kindergarten survey and kindergarten administrative panel data on service characteristics. To our knowledge, our study is the first to consider both individual and kindergarten traits over time as factors that drive satisfaction with kindergartens. Our sample consists of 410,162 individuals across 4113 kindergartens and 10,554 kindergarten years. As pointed out by Van Ryzin (2011), one reason for the debates over the relationship between government performance and user satisfaction is that few studies have combined subjective data with objective data about specific service providers. Our study fills this gap by combining individual-level time-series survey data with kindergarten-level panel data measuring service quality.

This article proceeds as follows. After having sketched the kindergartens' role in the Norwegian educational system, we present our arguments and expectations. We then go on to give an overview of the data sources and the methods used. Finally, using different multilevel modeling techniques, we test our arguments. We conclude by offering suggestions for further research.

19.1.1 Norwegian Kindergartens: Institutional Setting

The kindergarten institution in Norway is seen as a fundament for the Norwegian education system (Meld. St. 21 (2016–2017), 2017). The sector has seen major changes over the last decades concerning the kindergartens' framework conditions, reorganization and the owner structure. Kindergartens have existed since the late 1800s, but the scope was small. During the 1970s the modern kindergarten childhood was introduced and developed, creating a transition from children belonging exclusively in the home and their neighborhood but also in an institutional community of children outside of the home. The Kindergarten Act was introduced in 1975 with the objective to; "(…) *secure all children good opportunities for development and activity in close understanding and cooperation with the child's home* (§1)" (Ot.prp. Nr. 23(1974–75)). The idea of the kindergarten as a pedagogical institution benefiting all children through a specific practice of profession was there from the beginning (Korsvold, 2005).

The kindergarten as a welfare institution has, since the 1970s, gradually moved from being a selective and limited service to becoming a universal welfare service. A rapid development of the service came with the Kindergarten compromise in 2003, leading to a massive expansion of the kindergarten service, and a corresponding increase in the share of children enrolled in kindergartens and the full coverage of kindergartens in 2006. In 2009 all children were granted the right to a place in a kindergarten from the age of 1, manifesting the kindergarten as a universal welfare service. This expansion is reflected in the share of children enrolled in Norwegian kindergartens. In 1970, 2.8% of children below school age was enrolled in kindergartens. In 2000 the corresponding number was 62%, and in 2020 97% of all children above 3 years old are placed in kindergartens. For children under the age of 3 the corresponding number is 85% (SSB, 2020).

In parallel with the rapid expansion of kindergartens, the institution has increasingly been considered a vital part of the educational pathway in Norway. The most explicit manifestation of this change is seen in 2005 when the responsibility for kindergartens was moved from the Ministry of Children and Family Affairs to the Ministry of Education and Research. There is great focus on kindergarten as an area of intervention, where inequality can be fought at an early stage through the building of social competence, language learning and general learning through activities. All kindergartens are subject to the same national legislation, and although they are voluntary to use, they are widely seen as a first step on the educational ladder (Trætteberg & Lidén, 2018).

Despite being part of the social democratic welfare regime, the kindergarten institution's development in Norway has been marked by being mostly on private hands. In 2020, 53% of kindergartens were privately owned (SSB, 2020). Still, there are strong regulations on the private providers in terms of maximum fee (capped at a low level and thus heavily subsidized) and the pedagogical offer. Nevertheless, when it comes to the room for creating distinctive services, private providers do have flexibility (Børhaug et al., 2011; Børhaug & Lotsberg, 2012). The provider decides, for example, priority areas, admission requirements, opening hours and the number of staff beyond minimum requirements of pedagogical leaders (Kindergarten Act, 2005).

Historically, the private actors were often non-profit actors (ideal or voluntary organizations). After the kindergarten compromise in 2003, which was supported by a unanimous parliament, it was decided that all families that wanted their child in a kindergarten should have this opportunity. To achieve this, the authorities agreed that private and public kindergartens should be treated equally and different incentives such as public investment grants, full VAT compensation, attractive loans and cheap sites made available by the municipalities were used to have more private actors build and run kindergartens (Kunnskapsdepartementet, 2019, p. 9). The new incentive structures marked a new era for the private kindergartens, away from the small, non-profit private actors and toward bigger for-profit actors (Kunnskapsdepartementet, 2019, s.6). More and more kindergartens are bought and fused into owners with many kindergartens and a big central organization. The majority of these actors are driven by commercial companies that buy smaller kindergartens (Børhaug & Moen, 2014; Schade, 2018). Kindergartens are a municipal service, meaning that the municipalities are responsible for running public kindergartens and for the funding of private kindergartens. They are also the supervisory authority for all kindergartens in their municipality. The role of the municipality as the kindergarten authority is identified in the Kindergarten Act §8: *"the municipality is the local kindergarten authority"*. The municipality sees to it that the kindergartens are driven in accordance with current regulations.

There is free choice with kindergartens, meaning parents can apply to any kindergarten they want. However, due to capacity shortages, real user choice is limited in many municipalities. Central to our study the Kindergarten Act demand that there are pedagogues employed in all kindergartens. The act states that staff must be adequate so that the personnel can secure a satisfactory pedagogical operation. In 2020, 42% of the kindergarten personnel were preschool teachers, which implies you have a three-year education program at university college. The current manning norm implemented in 2018 holds that one must have one employee per three children when they are below three years old and one employee per six children when they are above three years old. In terms of pedagogic staffing, there must be one pedagogical leader for every seven children when they are below three years and one pedagogical leader for every 14 children when they are above 3 years old. Despite this being an improvement in the number of adults in the kindergarten, critique has been raised that this minimum requirement for staff is not fulfilled as the staffing varies a lot during the day and that it is only the core hours (10–14) that the norm is followed. Thus, staffing is the focus in our study.

19.2 Service Satisfaction and Quality: Expectations

Users' satisfaction with a particular welfare service may depend on a long list of factors. Experience with the services is just one factor. Media stories, information from family and friends, political preferences and demographic characteristics may also have an impact (Walle & Van Ryzin, 2011). Previous research has suggested a number of *user characteristics* as drivers of user satisfaction (Sitzia & Wood, 1997). An important insight is that the level of expectations toward welfare services affects individuals' level of satisfaction. The evaluation of services also depends on service *characteristics* (Bouckaert & Walle, 2003), such as frequency of use, homogeneity and heterogeneity of the services, directness of contact with the service and whether or not the service includes elements of professional judgments. Furthermore, the *local context* where the service is being delivered may influence individuals' attitudes and preferences. Thus, it should come as no surprise that research struggle to report a one-to-one relationship between service satisfaction and the quality of the services as such.

To study the service quality paradox in Norwegian kindergartens we build our expectations on this previous research and the “Disconfirmation of Expectation” (DOE) model (Oliver, 2010). Satisfaction studies have been criticized for the lack of a theoretical framework, and the DOE model is a response to this criticism and has become dominant in the study of user satisfaction in the public sector (James, 2007, 2011; Walle & Bouckaert, 2007). The premise in the DOE model is that citizens have expectations about services based on prior experiences, information from different sources (such as family, friends or the media) and preferences for provision of services. These expectations are the standard that they use to evaluate services. If there is no difference between expectations and performance, they will turn out to be satisfied. Satisfaction then does not imply high-quality services but rather that they are acceptable. If performance exceeds users’ expectations (“positive disconfirmation”) or if it falls short of their expectations (“negative disconfirmation”) the models state that high performance leads to positive disconfirmation while high expectations result in negative disconfirmation. Thus, satisfaction is not seen as a pure reflection of quality but as a combined result of both cognitive and attitudinal factors. While previous satisfaction models would predict that users’ satisfaction with their kindergarten will depend on the (objective) quality of the services provided, the DOE model claims that satisfaction will depend not only on the expectations users have but also on the evaluation of the actual services.

Previous research has found a relationship between service characteristics and satisfaction in the health sector (see for instance Wendt et al., 2010). Our argument is that this should also be the case when it comes to Norwegian kindergartens. That is, the quality of the services provided by the kindergarten should have an impact on users’ satisfaction with their respective local kindergarten. One crucial (and debated) aspect of the local kindergarten is staffing, and that is also the focus in our empirical analysis. The register data we have access to makes it possible for us to single out who and how many that spend their working days with the children and how this “child-to-staff” ratio changes over time across kindergartens. We expect that satisfaction with staffing will depend on the number of children per adult. The fewer children per adult the more satisfied we expect the children’s parents to be. In addition, we predict distinct differences in parents’ attitudes toward kindergarten staffing depending on the age of their child. Given the manning norm (see above) more staff is required for young children (<3 years of age) compared to older ones (>3). The argument behind the manning norm is that young children have greater care needs, and we therefore also expect parents with a young child to be more sensitive to variation in staffing. On the other hand, kindergartens may choose to prioritize staffing among the smallest children and low staffing may therefore to a larger degree “hurt” the oldest children and impact parents’ satisfaction with the staffing in the kindergarten. Before presenting our results, we give an account of our data, variables and estimation strategy.

19.3 Data and Variables

Our dataset consists of individual kindergarten-level data from the *Norwegian Kindergarten Parent Survey (NKPS)*, and kindergarten-level data from the *administrative register of Norwegian kindergartens (Norwegian abbreviation BASIL).* The NKPS is an annual survey administrated by The Norwegian Directorate for Education and Training (UDIR), carried out since 2016. The survey is distributed electronically and fielded between November 1 and December 20 each year. We utilize the first four waves (2016–2019). Approximately 100,000 parents have participated in each of the four years, and our analytic sample consists of 419,162 individual respondents. To secure balanced panel data at the group level, we only include kindergartens who participated all four years, resulting in a sample of 4,113 kindergartens. BASIL contain information on a range of indicators, such as the number of employees, employees' education and gender, number of children enrolled, size of outdoor and indoor areas, opening hours, ownership, pedagogical profile, and much more. Crucial to our study it also includes service characteristics such as staffing, public and private ownership, size, and department organization. This BASIL data is collected yearly through forms filled out by the kindergarten manager and is among other things the basis of allocation of funds and personnel from the municipality.

Each kindergarten decides whether to send the NKPS to its parents, and the survey can therefore not be considered representative in the same way as a probability-sample survey. However, the kindergarten level data from *BASIL* allows us to assess validity by comparing the kindergartens participating in the survey with the total population of kindergartens (see Appendix Table 19.4). The BASIL register shows that Norway has around 6000 kindergartens, of which two-thirds—around 4000—have participated in all four surveys. In addition, when we divide the number of parent respondents with the total number of children registered in BASIL, we find that approximately 40% of children in Norwegian kindergartens are represented in the parent survey each year. UDIR reports that nonresponse analyses show a 70% response rate among parents who received an invitation to participate; this share is stable over the survey waves (Norwegian Directorate for Education & Training, 2021). As the survey contains no socioeconomic and demographic data on parents, we have limited opportunities to assess whether certain groups are under- or overrepresented among the respondents. We can, however, compare the share of respondents reporting a minority language with the share of children registered with as minority language speakers at the kindergarten level. This comparison indicates that parents of children speaking a minority language are not underrepresented—contrary to the common tendency in survey research. The mean scores and distributions of our independent variables at the kindergarten level are compared between our study sample and the population (see Table 19.4). To assess whether potential differences are due to random variation, we have performed *t*-tests. We find that our included kindergartens have a higher mean child-to-staff ratio compared to the population (difference: 0.1); they have a higher mean number of children enrolled (difference: 9 children); and they have a higher share of employees with pedagogical education (difference: 2 percentage points). These differences are statistically significant. In addition, chi-square tests show an overrepresentation of municipally owned kindergartens (difference: 5 percentage points) and underrepresentation of kindergartens owned by a proprietorship (difference: 6 percentage points). There were no significant differences in the distribution of kindergartens with different department organizations. The high rate of representation, solid response rate and the modest differences between our study sample and the population lead us to argue that the parent survey is well-suited in order to study user satisfaction. We have merged the two data sources using kindergarten ID codes. The result is a dataset that allows us to study the relationship with 419,162 users' satisfaction with staffing and actual staffing in a diverse set of 4113 kindergartens.

The NKP survey includes a total of 32 questions where parents are asked to assess different parts of their child's kindergarten, such as indoor and outdoor area and the child's well-being and development. They are also asked to report the child's age, gender and language. In the parent survey, a range of variables measures satisfaction. The primary aim in this chapter is to compare subjective and objective quality indicators, and we have chosen satisfaction with staffing as the dependent variable in the analysis. This is the individual-level variable with the most evident "objective equivalent" at the institutional level. The dependent variable is parents' response to the following statement in the NKP; *"I experience the staffing density – the number of children per adult – as satisfactory"*. The item uses a Likert scale with five response categories ranging from "strongly agree, agree, neither agree nor disagree, disagree, to strongly disagree" in addition to a "don't know" response. In the analysis, we remove the "don't know" responses (1.6% of our raw sample). For the sake of simplicity, and to ease the interpretation of the results we have recoded and standardized the original response scale to a simple scale ranging from 0 to 100. Thus, the regression coefficients can be interpreted as percentage point changes in satisfaction. The mean score of the dependent variable is at approximately 75, and over three quarters of the respondents agree or strongly agree with the statement. This means that we analyze variation between a highly satisfied sample of parents.

The NKP survey only includes individual-level variables related to the child. Gender is a dichotomous variable with female as the reference category. The descriptive statistics in Table 19.1 show that around 52% of the respondents have boys. Language is a dichotomous variable where Norwegian, Swedish, Danish, English and Sami language is the reference category, and other mother tongue has the value 1. The latter category represents 21% of the respondents. Age is a categorical variable coded into four years or older and three years or younger. This dichotomization is done on the basis of the staffing norm of 2018, where different norms apply for children above and below three years of age. 58% of the respondents have children in the youngest category. This variable will also effectively control for compositional age effects between kindergartens (e.g., that some kindergartens may have a higher share of older children).

Turning to the *kindergarten-level* data (BASIL) our key variable of interest is staffing. We have calculated staffing as a child-to-staff ratio in which the total number of children is divided by the number of employees in relevant categories. The rich

Table 19.1 Descriptive statistics, individual-level variables ($N = 419{,}162$)

	Mean or %	Std. dev	Min	Max
Satisfaction with staffing	74.77	30.13	0	100
Gender				
Female	48.22			
Male	51.78			
Mother tongue				
No/Sw/Da/Sami/Eng	79.35			
Other	20.65			
Age				
4 years or older	41.64			
3 years or younger	58.36			

register data enables us to single out who actually spend their working days with the children. From this follows an inverse relationship between the indicator and the "quality" of staffing, but the calculation is done in accordance with the official way of assessing kindergarten staffing, and with the wording in the dependent variable. The average staffing of the kindergartens in our sample is 3.3 children per adult. When we stratify staffing by survey wave, we find that this estimate has decreased from 3.4 in 2016 to 3.2 in 2019.

The BASIL data include a wide set of control variables that could be relevant for parents' satisfaction. We include the number of children enrolled in the kindergarten (size). It has been argued that smaller classes may increase student performance (Leuven et al., 2008), and this may be the case for parents' satisfaction with kindergartens as well. This variable is divided by 10 to ease interpretation. The mean score of this variable is 5.6 for our analytic sample (i.e., 56 children). We also include the share of staff with pedagogical education which can be relevant for how parents experience service quality. The variable is calculated by adding up the number of employees with pedagogical education at tertiary or vocational level, divided by the number of employees in relevant categories (as calculated above). This ratio varies from 0 to 1 and is multiplied by 10 to ease interpretation, it thus indicates 10% differences or changes. The mean score of this variable is 0.62, meaning that 62% of employees have pedagogical education, while the rest have other educational backgrounds (at tertiary or vocational level) or are formally unqualified. Furthermore we include organization as a dichotomous control variable indicating whether the kindergartens organize the children groups in traditional, separate departments (reference), where each group has an assigned space, or as open-group centers, where children groups and common areas are larger (see Bratterud et al., 2012; Skalickà et al., 2015). Table 19.2 shows that 11% of kindergartens in our sample are open-group centers. Lastly, kindergarten ownership is measured by a categorical variable. The categories are public, i.e., municipality ownership (reference), joint-stock companies, non-profit

Table 19.2 Descriptive statistics, kindergarten-level variables ($N = 4113$)

	Mean or %	Std. dev	Min	Max
Staffing (child-per-staff ratio)	3.25	0.56	0.76	6.09
Staff's education (share)	0.62	0.16	0	1
Size (*N* children/10)	5.60	3.10	0.20	38.31
Organization (%)				
Department	88.72			
Open-group	11.28			
Ownership (%)				
Other ownership	4.64			
Joint-stock company	26.67			
Foundation	4.96			
Cooperative	11.21			
Proprietorship	1.00			
Municipality	51.52			

foundations and cooperatives, proprietorship, and other ownership (a diverse and relatively small category including kindergartens owned by housing associations, church councils, and more). Previous research indicates that parents in private kindergartens are somewhat more satisfied compared to parents in public kindergartens (Lindén et al., 2017; Trætteberg & Fladmoe, 2020). Descriptive statistics show that municipally owned kindergartens make up around half of the sample (52%). Joint-stock companies (27%) and cooperatives (11%) are also common owners. All variables included in the analysis together with their definitions are presented in Tables 19.1 and 19.2. For the group-level variables, the means and shares are calculated based on the included kindergartens, not on the distribution of respondents in kindergartens.

19.4 Statistical Approach

A common way of conducting multilevel survey research in the social sciences has been through cross-sectional survey data at the individual level combined with administrative data at a higher group level, often geographical units like municipalities, regions, and countries. In these analyses, the parameters are traditionally fixed at the individual level and random at the group level. As high-quality cross-national surveys have become available in multiple waves, for instance, the World Value Survey (seven rounds since 1981), the European Social Survey (nine rounds since 2002), and the Norwegian Citizen Survey (six rounds since 2009), researchers have been able to utilize the panel structure at the group level. By utilizing a variable measuring time, this data can be structured into a group-year level in addition to the individual and group levels. We note that this is the case with our data as well.

Schmidt-Catran and Fairbrother (2016) show that there are numerous ways to approach and model data with this structure. The coefficients of the group-level variables can be estimated through random effects (RE) models which use a weighted mean of between- and within-unit effects. Using an RE model in the context of this chapter would imply that we could not interpret whether the size and direction of the staffing variable coefficient indicates differences (also unmeasured) between different kindergartens, or whether it indicates the effect of a kindergarten changing its staffing from one year to the next. If the group-level data has a panel structure, fixed effects (FE) modelling, which estimates coefficients using only within-unit variation, is an alternative. This can for example be done by including the level-2 identifier as a dummy variable. The FE approach controls for time-invariant variation between units and is often presented as a solution to issues with confounding associated with RE models (Bell et al., 2019). However, FE has proved to be unreliable when the number of level-2 units is small, and the theoretical and practical interpretations are not always transparent; by using an FE model, you implicitly only investigate time-variant relationships (Mummolo & Peterson, 2018).

Fairbrother (2014), and later Bell et al. (2019), present a modelling strategy which we will follow in this chapter. This strategy builds on the strength of both the FE and RE modelling and has been conceptualized as a within-between random effects model (often abbreviated REWB)—sometimes termed a hybrid model. In this approach, we decompose the continuous and time-variant group-level variables. First, we construct a variable used to estimate the between-unit effect, i.e., a cross-sectional association with the dependent variable. We do this by calculating the mean score of each unit, thus creating a variable which only varies between units and not over time. In this case, this is each kindergarten's mean staffing across the four survey waves. Second, we construct a variable which estimates the within-unit effect, i.e., a longitudinal association. This is done by centering the variable, meaning we subtract each observation for each unit from the mean. Similar to a fixed-effects model, the within-effect variable here indicates how much each observation deviates from each unit's "own" mean, i.e., how the staffing indicator from 2016, 2017, 2018 and 2019 deviates from the mean staffing across all survey waves. If a kindergarten does not change its staffing over the survey years, the within-effect variable will have the value 0. Bell et al. (2019) argue that FE models often are used uncritically, as a technical solution to avoid possible endogeneity in RE estimates, without consideration of the substantial consequences of choosing an FE approach. They further argue that the REWB model allows a wider range of research questions, as FE models limit us to time-variant variables.

Consequently, our models assume a hierarchical clustering in four levels: the municipality level, the kindergarten level, the kindergarten-year level and the parent level. The analysis is performed through four steps. We start fitting a baseline model (also called "empty", "unconditional" or "null" model) without independent variables. This is used as a reference to the later models, and to calculate the size and significance of the intraclass coefficient (ICC), the share of variance distributed at our four levels. Next, we fit a bivariate model which includes the two version of the staffing variable: within-unit and between units. This model also includes time fixed effects, a survey wave dummy variable, which controls for temporal trends or events that have equal impact on all units in the sample. Then we proceed to fit a multivariate model, which includes all independent variables at the individual and institutional level described above. Finally, we fit models 4 and 5 which is based on model 3, but in addition include cross-level interaction terms between the child's age at the individual level and staffing at the kindergarten level. We fit separate interaction models for the within- and between-unit variables.

19.5 Results

We start by asking if there is significant variation in satisfaction with staffing at the different levels. The first column in Table 19.3 contains the results from the baseline model. The ICC scores tell us how much of the overall variation can be attributed to the different levels. Most of the variation is located at the individual level (as is usual for such data) and least at the municipal level. 1.6% of parents' satisfaction with staffing varies between municipalities, 9.2% between kindergartens nested in municipalities and 12.7% between kindergartens over time. The residual variance, 87.3%, is located at the individual level. This distribution of variance suggests that a substantial part of parents' satisfaction with staffing in their children's kindergartens can be explained by factors at the kindergarten level, and that multilevel modelling thus is a reasonable approach.

Next, we proceed to include the staffing variable from the kindergarten registers. Model 2, in the second column, shows that the effect is in the expected negative direction and significant. This means that the higher number of children per staff, i.e., poorer objective staffing, the lower is the parents' subjective satisfaction with kindergarten staffing. This holds for both components of the variable, meaning (1) that kindergartens with a higher mean child-to-staff ratio over the survey years have approximately 3.1 percentage points less satisfied parents (between variable) and (2) that kindergartens which improves their staffing over the survey years are predicted to experience a 2.5 percentage point increase in parents' satisfaction (within variable). The between variable ranges from 0.8 and 6.0, implying that the difference in predicted satisfaction between the kindergarten with highest and lowest average staffing is at around 16.5 percentage points. The within variable varies between −1.9 and 2.0, meaning that the kindergarten improving its staffing the most is predicted to have 9.8 percentage points higher satisfaction than the kindergarten with the largest decline. Model 2 also controls for a time trend by including estimates from the survey wave dummy variable, showing that there is a negative trend in parents' satisfaction with kindergarten staffing. For instance, a parent is predicted to have 6.5 percentage points lower satisfaction in 2019 compared with the first wave in 2016.

We now turn to the multivariate model. Model 3 shows that having a boy child has a significant negative association with staffing satisfaction. However, this is a weak and hardly substantially significant relationship. Parents of girls are less than one percentage point more satisfied. Furthermore, we find that the child's language has a

Table 19.3 Regression coefficients, DV: satisfaction with staffing

	Model 1	Model 2	Model 3	Model 4	Model 5
Staffing (between)		−3.143***	−2.571***	−1.901***	−2.571***
Staffing (within)		−2.515***	−2.317***	−2.318***	−2.343***
Year: 2017		−2.229***	−2.273***	−2.273***	−2.273***
Year: 2018		−5.835***	−5.939***	−5.944***	−5.939***
Year: 2019		−6.480***	−6.703***	−6.709***	−6.703***
Gender: Male			−0.907***	−0.907***	−0.907***
Language: Other			7.298***	7.299***	7.298***
Age: 3 years or younger			1.548***	5.668***	1.548***
Size (between)			−0.802***	−0.798***	−0.802***
Size (within)			−0.552***	−0.552***	−0.552***
Pedagogical education (between)			0.383***	0.382***	0.383***
Pedagogical education (within)			0.019	0.020	0.019
Organization: Open-group centre			−0.195	−0.194	−0.195
Ownership: Other			10.63***	10.64***	10.63***
Ownership: Joint-stock company			6.049***	6.055***	6.049***
Ownership: Foundation			9.513***	9.523***	9.513***
Ownership: Cooperative			10.99***	10.99***	10.99***
Ownership: Proprietorship			12.22***	12.18***	12.22***
Interaction: Age # Staffing (b)				−1.214***	
Interaction: Age # Staffing (w)					0.0457
Constant	76.48***	90.58***	85.65***	83.36***	85.65***
Variance: Municipal level	15.18***	13.12***	12.43***	12.47***	12.43***
Variance: Kindergarten level	70.25***	71.39***	50.07***	50.01***	50.07***
Variance: Kindergarten-year level	31.47***	23.29***	22.86***	22.87***	22.86***
Variance: Individual level	806.1***	806.2***	797.7***	797.7***	797.7***
N, municipality	374	374	374	374	374
N, kindergarten	4113	4113	4113	4113	4113
N, kindergarten-year	10,554	10,554	10,554	10,554	10,554
N, individual	419,162	419,162	419,162	419,162	419,162

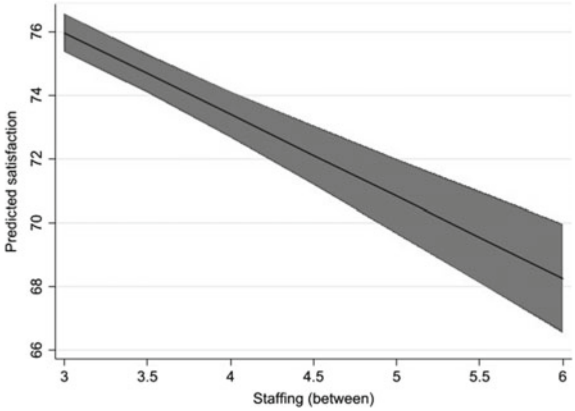
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

strong, positive and significant association with parents staffing satisfaction; having a child whose mother tongue is not Norwegian, Swedish, Danish, English or Sami is associated with around 7 percentage points higher satisfaction. This is an interesting finding and may indicate that kindergarten is an important arena for integration of children of immigrants. This is certainly a topic in need of further research to single out the mechanism underlying this effect. The dichotomized variable for the child's age shows that parents of children aged three years or below are around 1.5 percentage points more satisfied with kindergarten staffing.

Turning to the kindergarten level we briefly look at the control variables at that level. The between- and within-variables measuring kindergarten size has weak, negative associations with the dependent variable; satisfaction is 0.8 percentage points higher for every 10 children in mean size (between), and increases by 0.6 percentage points when the number of children increases by 10 (within). For the variable measuring staff's education, the between-unit estimate is statistically significant, and indicates that in kindergartens with 10% higher mean share of staff with pedagogical education, parents' satisfaction is 0.4 percentage points higher. Our results do not indicate that department organization is relevant for parents' satisfaction with staffing; this variable is not significant. Ownership appears to have strong associations with parents' staffing satisfaction. Here, the municipally owned kindergartens score significantly lower than all other categories, with the largest gap—12.2 percentage points—to kindergartens owned by proprietorships. Pairwise comparisons of the different private ownership categories (not reported) indicate that there is a tripart division: parents in joint-stock company-owned kindergartens are significantly less satisfied than parents in kindergartens owned by non-profit organizations (foundations, cooperatives, proprietorships and other ownership).

More important, and turning to our key variable of interest the associations between subjective and objective staffing is significant also when controlling for other covariates. The between estimate is reduced to −2.6 and the within estimate to −2.3 in the multivariate model. Figures 19.1 and 19.2 illustrate these associations

Fig. 19.1 Predicted satisfaction and staffing (between units)



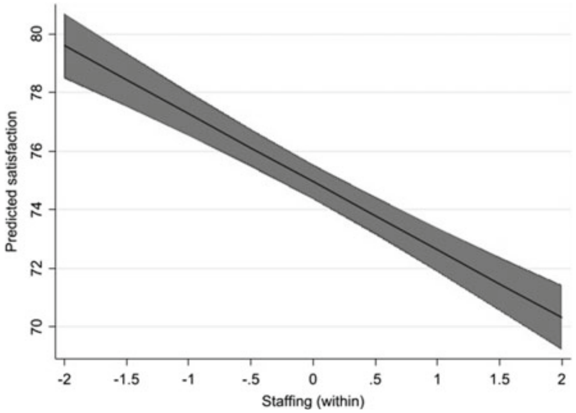


Fig. 19.2 Predicted satisfaction and staffing (within units)

through predicted satisfaction scores. In Fig. 19.1, we plot selected values of the between variable: the *x*-axis varies between the staffing norm for children per adult obliged by law in 2018: 3 children under 3 years and 6 children of 4 years and older. We can observe that parents in kindergartens with the highest mean child-to-staff ratio are predicted to be approximately 9 percentage points less satisfied than parents in kindergartens with the lowest mean ratio.

Figure 19.2 shows the effect of a kindergarten increasing or decreasing the number of children per adult. The *x*-axis is scaled to fit the range of the within variable, and we see that kindergartens that *decrease* their child-to-staff ratio by 2 children have 10 percentage points higher parent satisfaction compared to kindergartens that have *increased* their child-to-staff ratio by 2 children. As the width of the 95% confidence intervals for this variable reflect, over 90% of the kindergartens vary between −0.5 and 0.5; the predicted difference in satisfaction between these two values is around 2 percentage points.

Finally, models 4 and 5 in Table 19.3 include the interaction terms between the age of the child and kindergarten staffing. The results reveal that only the between-unit interaction returns a statistically significant estimate. When the oldest age category is a reference, a negative estimate of −1.2 in model 5 implies that the association between subjective and objective staffing is stronger for parents with young children. This is as we expected. As the age variable has a positive estimate, meaning that parents of the younger children are more satisfied, the direction of the interaction effect suggests that the satisfaction difference between age groups is smaller in kindergartens with poor staffing. Figure 19.3 illustrates the marginal predictions of this interaction: The legend for parents of the youngest children (dotted line) has a steeper slope. In kindergartens with a child-to-staff ratio at 3, parents of the younger children are approximately 2 percentage points more satisfied with staffing, but in kindergartens with a ratio above 4, the 95% confidence intervals overlap and there is no significant difference in satisfaction between the age groups.

Turning to the variance components there are numerous ways to calculate the reduction in variance across multilevel models. A relative straightforward way which

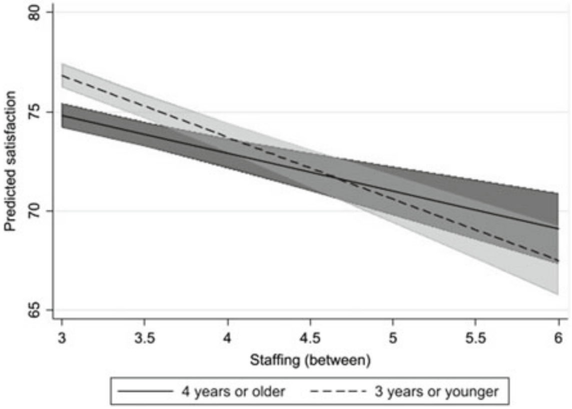


Fig. 19.3 Predicted satisfaction and interaction age # staffing (between units)

provide intuitive results is to estimate the relative change in residual variance between the baseline and full models. A drawback of this method can be that in some cases, the unexplained variance may increase after variables are added, as we can observe at the individual and kindergarten level between the baseline and bivariate models. This may occur if a variable at a lower level (kindergarten-year) is introduced, without the between-group (kindergarten) variance being affected (LaHuis et al., 2014). When we compare the baseline model (Model 1) to the multivariate model (Model 3), we find that 1% of variance at the individual level is reduced by the variables included in our model. 27% of residual variance at the kindergarten-year level in the baseline model is reduced in the multivariate model. At the kindergarten level, we find a 29% reduction. At the municipality level, 18% of residual variance is reduced; since we do not include any variables at this level, the reduction is most likely due to different composition of kindergartens within municipalities.

19.6 Conclusion

The service quality paradox indicates that research struggle to report a clear-cut relationship between how service quality at the institutional level and changes in such services impact citizens' satisfaction at the individual level. This study contributes to research on user satisfaction, by focusing on the impact of kindergarten quality measures and changes in such quality measures.

Few studies have access to data measuring both subjective evaluations and objective performance which is vital to unravel the connection between welfare state inputs and outputs. Based on four waves of a large survey among Norwegian parents, we find that parents' satisfaction with kindergarten staffing varies considerably not only among parents, but between kindergartens as well. This result alone means that it is important to study institutional determinants when trying to disentangle different aspects of user satisfaction. We show that features of service provision at the kindergarten level matter to parents' evaluation of staffing local kindergartens. Crucial to our argument, we find that fewer children per adult are associated with increased satisfaction with kindergarten staffing. We also find that parents with the youngest children are more sensitive to the number of children per adult. Thus, our study supports research that has found a link between aspects of service quality/provision and user satisfaction.

Many of the research questions in the study of the welfare state are multilevel in nature and require theoretical renewal, but also methodological innovation. Multilevel techniques which distinguish between both variation and causation at different levels are well-suited to address important issues in the welfare state. Future research should dig deeper into how different types of local service institutions and reforms in these institutions impact not only attitudes, but also behavior among their users. This will provide more insights into how the welfare state works.

Appendix

See Table 19.4.

Table 19.4. Descriptive statistics, kindergarten-level variables, analytic sample and population compared

	Analytic sample		Population	
	Mean or %	*N*	Mean or %	*N*
Staffing (child-per-staff ratio)***	3.25	4113	3.15	6094
Staff's education (share)***	0.62	4113	0.60	6218
Size (*N* children/10)***	5.60	4113	4.78	6094
Organization (%)		4113		5503
Department	88.72		87.82	
Open-group	11.28		12.18	
Ownership (%)***		4113		6223
Other ownership	4.64		6.57	
Joint-stock company	26.67		25.28	
Foundation	4.96		4.44	
Cooperative	11.21		9.88	
Proprietorship	1.00		7.04	
Municipality	51.52		46.79	

Note *** $p < 0.001$ in chi-square or t-test

References

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality & Quantity, 53*(2), 1051–1074. https://doi.org/10.1007/s11135-018-0802-x

Bouckaert, G., & Van de Walle, S. (2003). Comparing measures of citizen trust and user satisfaction as indicators of 'good governance': Difficulties in linking trust and satisfaction indicators. *International Review of Administrative Sciences, 69*(3), 329–343.

Børhaug, K., & Moen, K. H. (2014). *Politisk-administrative rammer for barnehageledelse.* Universitetsforlaget.

Børhaug, K., Helgøy, I., Homme, A., Lostberg, D. Ø., & Lundvigsen, K. (2011). *Styring, organisering og ledelse i barnehagen.* Fagbokforlaget.

Børhaug, K., & Lotsberg, D. Ø. (2012). Institusjonelle betingelser for konkurranse mellom offentlige og private tjenesteytere i barnehagesektoren. *Nordiske Organisasjonsstudier, 14*(2), 27–48.

Bratterud, Å., Sandseter, E. B. H., & Seland, M. (2012). *Barns trivsel og medvirkning i barnehagen: Barn, foreldre og ansattes perspektiver*. NTNU samfunnsforskning, Barnevernets utviklingssenter.

Easton, D. (1965). *A systems analysis of political life.* John Wiley.

Fairbrother, M. (2014). Two multilevel modeling techniques for analyzing comparative longitudinal survey datasets. *Political Science Research and Methods, 2*(1), 119–140. https://doi.org/10.1017/psrm.2013.24

Friman, M., & Fellesson, M. (2009). Service supply and customer satisfaction in public transportation: The quality paradox. *Journal of Public Transportation, 12*(4), 4.

James, O. (2007). Evaluating the expectations disconfirmation and expectations anchoring approaches to citizen satisfaction with local public services. *Journal of Public Administration Research and Theory, 19*, 107–123. https://doi.org/10.1093/jopart/mum034

James, O. (2011). Managing citizens' expectation of public service performance: Evidence from observation and experimentation in local government. *Public Administration, 89*, 1419–1435. https://doi.org/10.1111/j.1467-9299.2011.01962.x

Kindergarten Act [Lov om barnehager]. (2005). *LOV-2005-06-17-64.* Kunnskapsdepartementet.

Korsvold, T. (2005). *For alle barn!: Barnehagens framvekst i velferdsstaten (2. utg.).* Abstrakt forlag.

Kunnskapsdepartementet. (2019). Forslag til endringer i barnehageloven med forskrifter [Høringsutkast]. Hentet fra Høring av forslag til endringer i barnehageloven med forskrifter (ny regulering av private barnehager) - regjeringen.no

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17*(4), 433–451. https://doi.org/10.1177/1094428114541701

Meld. St. 21(2016–2017). (2017). *Lærelyst – tidlig innsats og kvalitet i skolen.*

Lindén, T. S., Fladmoe, A., & Christensen, D. A. (2017). Does the type of service provider affect user satisfaction? Public, for-profit and nonprofit kindergartens, schools and nursing homes in Norway. In K. H. Sivesind & J. Saglie (Eds.), *Promoting active citizenship: Markets and choice in Scandinavian welfare* (pp. 261–284). Springer International Publishing.

Mummolo, J., & Peterson, E. (2018). Improving the interpretation of fixed effects regression results. *Political Science Research and Methods, 6*(4), 829–835. https://doi.org/10.1017/psrm.2017.44

Norwegian Directorate for Education and Training. (2021). *Foreldreundersøkelsen i barnehage – deltakelse og svarprosent*. https://www.udir.no/tall-og-forskning/statistikk/statistikk-barnehage/foreldreundersokelsen-i-barnehage--deltakelse-og-svarprosent/. Accessed 28 May 2021.

Leuven, E., Oosterbeek, H., & Rønning, M. (2008). Quasi-experimental estimates of the effect of class size on achievement in Norway. *Scandinavian Journal of Economics, 110*(4), 663–693.

Oliver, R. L. (2010). *Satisfaction: A behavioral perspective on the consumer*. M.E. Sharp Inc. https://doi.org/10.4324/9781315700892

Ot.prp. Nr. 23(1974–75). *Lov om barnehager m. v.* Forbruker-og administrasjonsdepartementet.

Skalická, V., Belsky, J., Stenseng, F., & Wichstrøm, L. (2015). Preschool-age problem behavior and teacher–child conflict in school: Direct and moderation effects by preschool organization. *Child Development, 86*(3), 955–964. https://doi.org/10.1111/cdev.12350

SSB. (2020) Fakta om barnehager. Hentet fra Barnehager (ssb.no)

Schade, S. M. (2018). Barnehage og sykehjem – omfang, utvikling og organisasjonsformer. In I. B. Jensen, S. Antonsen, A. B. Erichsen, A. Berge, F. Voldnes, S. M. Schade & G. Høin (Eds.), *Kommersialisering av fellesgodene. Virkninger på skatteinntekter, lønn og samfunnsøkonomiske kostnader.* Høgskoen i Innlandet. Retrieved from https://brage.inn.no/inn-xmlui/handle/11250/2570157

Schmidt-Catran, A. W., & Fairbrother, M. (2016). The random effects in multilevel models: Getting them wrong and getting them right. *European Sociological Review, 32*(1), 23–38. https://doi.org/10.1093/esr/jcv090

Sitzia, J., & Wood, N. (1997). Patient satisfaction: A review of issues and concepts. *Social Science & Medicine, 45*(12), 1829–1843.

Trætteberg, H. S., & Fladmoe, A. (2020). Quality differences of public, for-profit and nonprofit providers in Scandinavian welfare? User satisfaction in Kindergartens. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *31*(1), 153–167. https://doi.org/10.1007/s11266-019-00169-6

Trætteberg, H. S., & Lidén, H. (2018). *Evaluering av moderasjonsordningene for barnehagen.* Delrapport 1. Rapport 2018:4. Institutt for samfunnsforskning.

Van de Walle, S. (2018). Explaining citizen satisfaction and dissatisfaction with public services. In *The Palgrave handbook of public administration and management in Europe* (pp. 227–241). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-55269-3_11

Van de Walle, S., & Bouckaert, G. (2007). Perceptions of productivity and performance in Europe and the USA. *International Journal of Public Administration, 30*, 1–18. https://doi.org/10.1080/01900690701225309

Van Ryzin, G. G. (2011). An experimental test of the expectancy-disconfirmation theory of citizen satisfaction. *Journal of Policy Analysis and Management, 32*, 597–614. https://doi.org/10.1002/pam.21702

Wendt, C., Kohl, J., & Pfeifer, M. (2010). How do Europeans perceive their healthcare system? Patterns of satisfaction and preference for state involvement in the field of healthcare. *European Sociological Review, 26*, 177–192. https://doi.org/10.1093/esr/jcp014

Håvard Thorsen Rydland is a Senior Researcher at NORCE. He received his Ph.D. from the Norwegian University of Science and Technology, Department of Sociology and Political Science, in 2020. His work has mainly centred around social inequalities in health outcomes, utilizing Norwegian and European survey and register data. He has been published in social science and public health journals.

Åsta Dyrnes Nordø is a senior researcher at NORCE social sciences. She is specialized in political behavior and public opinion formation. She has broad competence in quantitative methods to develop and analyze both survey and macro data, focusing on causal inference. She has been affiliated with the survey infrastructure Digital Social Science Core Facility (DIGSSCORE) at the University of Bergen since the startup in 2013. Currently, she chairs the Political Behavior section in the Norwegian Citizen Panel.

Dag Arne Christensen is a Research Professor at NORCE. He is a political scientist from the University of Bergen (Comparative Politics). His work includes topics such as user satisfaction, political behavior, and field experiments on voter turnout. He has experience with survey data and register data, and different quantitative methods. He has published his research in a wide range of political science journals.

Chapter 20
Multilevel Modeling and Assessment of the Study-Relevant Knowledge of First-Year Students in a Master's Program in Business and Economics

Susanne Schmidt, Olga Zlatkin-Troitschanskaia, and Marie-Theres Nagel

Abstract National and international studies on the assessment of students' knowledge of business and economics focus primarily on bachelor's students. There are few empirical findings on the domain-specific knowledge of students entering master's programs in business and economics, two of the most popular programs world-wide. In particular, there is a lack of studies based on validated test instruments, as required by the International Standards for Educational and Psychological Testing (AERA et al. 2014). In this paper, we address the question to what extent the test for assessing business and economics knowledge, which has been comprehensively validated according to the AERA standards, and so far only been used in studies focusing on bachelor's programs, also allows for the valid measurement of the domain-specific knowledge of students at the beginning of their master's studies. We assume that the final bachelor grade as well as the number of and the grade achieved in attended and completed specialized courses are positively correlated with domain-specific knowledge in business and economics. More specifically, we expect a stronger correlation between business knowledge and the grade in and the number of business courses attended in the bachelor's program.

Keywords Multilevel modeling · International standards for educational and psychological testing, Structural equation model, Knowledge test

20.1 Introduction and Research Focus

Economics has been one of the most popular and most frequently attended courses of study for years, in Germany and internationally (OECD, 2017, 2020; Statistisches Bundesamt, 2018). There are a number of studies both on the characteristics of students' economic knowledge at the beginning of a bachelor's degree course and on the personal and contextual factors (e.g., prior knowledge, gender, intelligence, type of university, degree program, etc.) that influence economic knowledge (e.g.,

S. Schmidt (✉) · O. Zlatkin-Troitschanskaia · M.-T. Nagel
Johannes Gutenberg University, Mainz, Germany
e-mail: susanne.schmidt@uni-mainz.de

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3_20

Schlax et al., 2020; Zlatkin-Troitschanskaia et al., 2019). In contrast, little research has been conducted on the level of students' economic knowledge at the end of a bachelor's degree and over the course of a master's degree (e.g., Happ et al., 2019; Kraitzek et al., 2020; Tse & Tam, 2017). Therefore, little is known about the factors that determine or predict success in an advanced course of study, such as a graduate's degree program.

For admission to a master's program, universities generally rely on (formal) performance criteria from the bachelor's program. These include the final bachelor grade, which is composed summatively of the module grades achieved throughout the degree program. The bachelor's grade is considered a suitable predictor of success in the master's program; nonetheless, there is a lack of empirical studies researching whether and to what extent this holds true for the domain of economics (Happ et al., 2019). In addition, several studies showed domain-specific knowledge to be the best predictor of academic success in the bachelor's program (e.g., Dochy et al., 2002; Schlax et al., 2020), while similar analyses of the master's program are lacking.

Based on this prior research, it can be assumed that study success in a master's program, if operationalized as subject-specific knowledge in various content areas (e.g., finance or accounting), can be predicted from the level study-related prior knowledge before the beginning of the graduate course and, in particular, the attendance of courses in the corresponding content areas during the bachelor's studies (Happ et al., 2019; Kraitzek et al., 2020).

In terms of theory, our study follows a content-based as well as cognition-based structural model of economic knowledge (for details, see Zlatkin-Troitschanskaia et al., 2014). Based on Germany-wide curricular analyses as well as expert ratings, domain-specific knowledge is modeled as a complex, multifaceted construct that can be subdivided into different business and economics content areas (e.g., microeconomics and finance) as well as into three different cognitive levels (i.e., understanding, explicit application, and implicit application, for details, see Zlatkin-Troitschanskaia et al., 2014). Therefore, methodologically, we apply the multilevel approaches and latent analyses to investigate the research question of this paper: to what extent can master's students' domain-specific knowledge in the content areas of accounting and finance be predicted from courses in the same content areas attended during bachelor's studies, if other theoretically expected influencing variables (such as gender or socio-economic background) are included and controlled for in the multilevel analyses presented below.

Based on the learning theory by Helmke and Schrader (2001), we assume that the final bachelor grade as well as the number of and the grade achieved in completed relevant courses positively correlate with master's students' level of domain-specific knowledge in economics. More specifically, we expect a stronger correlation between knowledge in specific content areas of business (e.g., finance) and the grade in and the number of the corresponding business courses attended during bachelor's studies; the same applies to the number of and grades in economics courses and economic knowledge (see Sect. 20.2).

We used Confirmatory Factory Analyses (CFA) models to test for the internal structure of the latent knowledge constructs of the two business subareas, accounting and finance. With Multilevel Multiple Indicator and Multiple Causes (MMIMIC) models, relationships between knowledge and dummy-coded variables of the courses attended were tested, whereby the influences of students' key personal characteristics such as migration background, gender, etc., were included as predictors and control variables (see Sect. 20.4). After presenting the analysis results in Sect. 20.4, we conclude with a critical discussion of the multilevel-modeling approach to validly assess the domain-specific knowledge of students in a master's program in economics.

20.2 Theoretical Foundation and Hypotheses

In terms of the most relevant factors influencing academic success and the development of subject-specific knowledge at the beginning and over the course of the (bachelor's) program, in addition to students' personal characteristics and preconditions (e.g., gender, migration, and educational background), following the well-established learning model by Helmke and Schrader (2001), students' learning opportunities (e.g., type and number of courses attended) and learning potential (e.g., intelligence, prior knowledge) must also be taken into account (Helmke & Schrader, 2011).

Regarding the personal socio-biographical characteristics of students that may influence subject-specific knowledge, according to previous findings (e.g., Brückner et al., 2015b; Happ et al., 2019; Owen, 2012), where socio-cultural background and gender are of particular importance. In Germany, socio-cultural background is primarily divided into the presence or absence of a migration background. The country of origin of students' parents (Germany or another nationality, Bellin, 2009) as well as the main (spoken) language, i.e., whether students' grew up speaking German or another primary language, are often used as indicators for migration background (e.g., Happ et al., 2019). In this respect, it has already been shown in various fields of economics, e.g., macro- and microeconomics as well as financing, that students with a different main language or country of origin perform worse in tests than fellow students without a migration background (Broecke & Nicholls, 2007; Brückner et al., 2015a, 2015b; Förster et al., 2015; Happ et al., 2019; Zlatkin-Troitschanskaia et al., 2015; Zorlu, 2011). This effect has also been found in other countries, for example, in the USA (e.g., Borde, 2017).

On the relationship between gender and economic knowledge, so far, several studies using standardized tests or course grades have found an effect in favor of male students (for accounting, see, Duff, 2004; Fritsch et al., 2015; Gracia & Jenkins, 2010; for financing, see, Borde, 2017; Terry, 2002), although some studies, found no significant difference in expertise between the genders (Brahmasrene & Whitten, 2001; Byrne & Flood, 2008; Keef & Roush, 1997; Park & Hayes, 1994; Paver & Gammie, 2005).

Based on the existing research, the following hypotheses can be formulated with regard to socio-biographical factors:

H1a: Male students have a higher level of subject-specific knowledge in the field of accounting (finance) than female students.

H1b: Students with a migration background have a lower level of subject-specific knowledge in accounting (finance) than students without a migration background.

With regard to learning opportunities, a distinction can be made between subject-relevant university and pre-university learning opportunities. While university learning opportunities are often operationalized as lectures attended as part of a course of study, pre-university learning opportunities for economic knowledge can include a completed vocational training and/or a school-leaving certificate (Abitur) from a school with an economic focus.

In terms of previous knowledge, initial findings in the content areas of finance and accounting show a correlation between the performance of students in their second and third year of study and their respective performance in the previous year (Gracia & Jenkins, 2010). Also, students performed better in finance courses if they had previously taken advanced placement courses (preparatory courses with the content of basic university courses). Attending subject-relevant courses also has an effect on domain-specific knowledge in the area of economics (Zlatkin-Troitschanskaia et al., 2015): a positive correlation was found between attending economics courses and the relevant subject-specific knowledge. Accordingly, whether at least one course has already been attended in a content area would be more relevant for the development of knowledge in, for example, finance and accounting than the study semester.

The influence of university learning opportunities on knowledge acquisition was also found for various areas of business and economics in bachelor's program in Germany (e.g., for finance, see Förster et al., 2015; for economics, see Zlatkin-Troitschanskaia et al., 2015). However, these studies were inconclusive as to the extent to which the learning opportunities attended during the bachelor's program influence academic success in terms of subject-specific knowledge at the end of the bachelor's program or in the master's program. The second hypothesis to be investigated is therefore as follows:

H2: Attendance of at least one course in accounting (finance) is positively related to the level of subject-specific knowledge at the beginning of the master's program in accounting (finance).

Since pre-university learning opportunities can have a significant influence on previous domain-specific knowledge, which in turn is of central importance for the acquisition of knowledge over a course of study, these must also be taken into account as influencing factors. In Germany, one particularly important learning opportunity for the earlier development of economic knowledge outside the university is the systematic teaching of economic topics as part of general school education at schools with economic focus or vocational schools (e.g., Solga et al., 2014). In general schooling, economics content is usually only taught within the framework of social studies teaching units or advanced economics courses (Federal Institute for Vocational Education and Training, 2018; Oberrauch & Kaiser, 2019; Weber, 2002). There are only few findings on the connection between prior knowledge acquired in school-based learning opportunities and academic success in terms of subject-specific knowledge in higher education economics (Schlax et al., 2020).

A second, more comprehensively researched opportunity to acquire economic knowledge outside the university in Germany is offered in the form of vocational training, in which practical experience can be gained in addition to the acquisition of knowledge in vocational school lessons. Findings from economic knowledge tests show that practical experience acquired during and before university studies leads to better results in various economic test subareas, i.e., students with completed vocational training tend to have better knowledge in subjects such as accounting (Fritsch et al., 2015), finance (Förster et al., 2015), and economics (Zlatkin-Troitschanskaia et al., 2015) than students without vocational training. The third hypothesis regarding influences on students' knowledge at the end of their bachelor's or beginning of their master's studies is therefore as follows:

H3: Students who have attended a school with an economic focus and/or completed vocational training in business have a higher level of domain-specific knowledge in accounting (finance) than students who have attended a school without an economic focus and/or students without relevant vocational education.

20.3 The Assessment of Economic Knowledge: Study Design, Instruments, and Sampling

As national and international studies on students' acquisition of economic knowledge focus primarily on bachelor students (e.g., Kim & Lalancette, 2013; Schmidt et al., 2016; Zlatkin-Troitschanskaia et al., 2016), there are only few test-based findings on the knowledge in economics among bachelor graduates or entrants to a master's program in economics (Happ et al., 2019; Kraitzek et al., 2020).

Generally, there is still a lack of studies that measure academic success using validated domain-specific tests, since so far only a few valid instruments are available for measuring economic knowledge (Zlatkin-Troitschanskaia et al., 2019). The standardized and validated WiwiKom test used in this study makes it possible to draw conclusions about the level of economic knowledge of master's students and to examine the hypothesized relationships with various influencing variables.

The WiwiKom project for modeling and measuring competencies in business and economics was part of the research program *Modeling and Measuring Competencies* *in Higher Education* (KoKoHs), which was founded in 2012 by the Federal Ministry of Education and Research to develop reliable test procedures for validly measuring subject-specific knowledge. In WiwiKom, a standardized test ('WiwiKom test') to assess the business and economic knowledge of university students was developed and comprehensively validated (Zlatkin-Troitschanskaia et al., 2014, 2019) according to the 'Standards for Educational and Psychological Testing' (AERA et al., 2014). The test is based on a structural model of economic knowledge that differentiates between various content areas and cognitive levels (for details, Zlatkin-Troitschanskaia et al., 2014).

The WiwiKom test version used in this study consists of a total of 58 items, whereby 28 items were developed to assess knowledge in the areas of financing and accounting. These 28 test items are based on the German adaptation (Zlatkin-Troitschanskaia et al., 2014) of the *Examen General para el Egreso de la Licenciatura* (EGEL; CENEVAL, 2010).

To limit the test processing time to 45 min and to ensure the processing of all items at the same time, a booklet design was used. Thereby, the 58 items were divided into seven test versions with 24–28 items each, so that the completion of different test versions takes about the same time.

In addition to the knowledge test data, the students' personal characteristics including gender, main language, school-leaving grade, as well as previous economic education (vocational training or attendance of a school with economic focus) were surveyed. In terms of contextual variables, the current study semester attended courses in accounting and finance as well as the final grades of these courses were surveyed as further study-related data.

In this paper, we used data collected in the third survey in the WiwiKom project. The dataset consists of 1523 master's students of economics at 27 universities and 13 universities of applied sciences, of which 1492 plausibilized cases were included in the following analyses.

The participants were 52.5% male and 47.5% female, aged between 20 and 49 years (on average 25 years). Approximately 1401 (93.9%) students were studying in a master's program at the time of the survey, 62% were enrolled at a university, and 38% were at a university of applied sciences; 23% of the participants reported previous knowledge of economics in terms from attending a school with an economic focus and 28% reported attending a corresponding advanced course in economics. About one-fifth of the participants (21%) had completed a vocational training. The students also differed with regard to their origin: while the majority of 1087 (72.9%) participants stated that they did not have a migration background, 14% reported having a migration background and German as their family language, a similar number (11.5%) reported having a migration background with a different family language, and 0.9% of the participants (13 students) were exchange students (Table 20.1).

Table 20.1 Sample description

Attributes	$N = 1492$ (%)
Gender	
Female	708 (47.5%)
Male	783 (52.5%)
Degree	
Bachelor	83 (5.6%)
Master	1401 (93.9%)
Higher education institution	
University	922 (61.8%)
University of applied sciences	570 (38.2%)
Migration background	
Yes, main language German	215 (14.4%)
Yes, other main language	172 (11.5%)
No	1087 (72.9%)
Exchange student	13 (0.9%)
Completed vocational training	
Yes	299 (20.8%)
No	1190 (79.9%)
School with economic focus	
Yes	343 (23.0%)
No	1144 (76.7%)
Economics major in school	
Yes	422 (28.3%)
No	1059 (71.0%)
Age (in years)	
Mean	2.5
Standard deviation	2.3

20.4 Methods and Results of Structural Equation Models in the Multilevel Approach

20.4.1 The Multilevel Approach in a Structural Equation Model

Compared to one-dimensional analyses, more complex procedures are required for the analysis of multilevel structures to account for the specificity of nested data. Due to the nesting of students in universities, the observations of students of the same university are not independent of each other. Accordingly, the specific nature of multiple-nested structures is that the responses of students from one university are more similar than compared to the responses of respondents from other universities (Snijders & Bosker, 2012). This represents a prerequisite violation of the usual analysis procedures commonly used for correlation analyses (e.g., classical regression analysis) (Hox et al., 2017). Here, it must hold that the random errors of the observed values are uncorrelated, homoscedastic, and normally distributed (Byrne, 2012). This is generally not the case, due to the dependence of responses in nested data. Therefore, a statistical procedure is required that can estimate the structure of the residuals explicitly. This is achieved by using the multilevel option in Mplus (Muthén & Muthén, 1998–2017) and the Institution ID as cluster variable. Furthermore, we took into account that the economic knowledge measured with the WiwiKom test is a latent construct. Therefore, we did not use an aggregated knowledge score, such as a sum score, but instead we modeled the latent score and the relationships between the knowledge score and its predictors in one model, which is called a multiple indicators multiple causes (MIMIC) model (Muthén & Muthén, 1998–2017). By combining the MIMIC analyses with a multilevel approach, we conducted multilevel multiple indicators multiple causes, i.e. MMIMIC (Schmidt et al., 2016) models to test our research hypotheses H1 to H3 (see Sect. 20.2).

20.4.2 Confirmatory Factor Analysis (CFA)

20.4.2.1 One-Dimensional CFA Models for Each Content Area

Before we tested the hypothesis H1 to H3, we constructed Confirmatory Factor Analysis (CFA) models using Mplus 8 (Muthén & Muthén, 1998–2017) to test the internal construct structure. Only if the internal structure of both subareas, accounting and finance, is well represented by the WiwiKom test, conducting MMIMIC models for hypothesis testing will provide valuable results. We used the WLSMV estimator for categorical data and the analysis type complex option in the modeling to take into account the multilevel structure of the data. Using this type of analysis allows us to account for the standard errors and consider that the multilevel structure of the data, due to the students belonging to different higher education institutions. As students from various universities have been observed, these observations have been taken into account in the evaluation of the data. For model evaluation, the following fit criteria are utilized: The chi-square (χ^2) statistic with the related degrees of freedom (df) and the ratio between χ^2 and df, the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI). To evaluate the model fit on the item level, we interpreted the Item Response Theory parameters which are automatically generated in Mplus when conducting single-factor CFA models and enabled a more model-based interpretation.

First, to evaluate the internal structure and the construct validity of the test version used in this study, two confirmatory factor (CFA) models were estimated: one for accounting and one for finance. The results for the accounting area are presented in Table 20.2.

Table 20.2 CFA model fit accounting (Model 1.1 and Model 1.2)

Model	χ^2 (df)	p	χ^2/df	RMSEA [C.I.]	CFI	TLI
1.1: 1-factor accounting (16 items)	132.932 (104)	0.029	1.28	0.016 [0.006–0.024]	0.946	0.937
1.2: 1-factor accounting (14 items)	101.023 (77)	0.035	1.31	0.017 [0.005–0.026]	0.958	0.950

For accounting, the expected one-dimensional structure is well represented with the 16 corresponding test items (see Model 1.1 in Table 20.2). All fit criteria are close or beyond the cutoff values: the χ^2-statistic is rather insignificant and the ratio of χ^2 and degrees of freedom (df) is below 2. RMSEA is below 0.05, with a narrow confidence interval (C.I.). CFI is close to 0.95 and the TLI is close to 0.94. These criteria are all favorable and indicate a well-fitting model (Brown, 2015).

This became also evident regarding the local level (see Table 20.3). Almost all items are positively related to the factor. Only two items show a non-significant factor loading and threshold (LR6 und LR21) and one item shows only a non-significant threshold (LR5) (see Table 20.3). To test, whether these items are still meaningful regarding the construct definition and whether these items should be kept for further modeling and not eliminated from the model, another factor model was estimated without items LR6 and LR21 (see Model 1.2 in Table 20.2).

Table 20.3 IRT item parameters for Model 1 (accounting)

	Factor loading (item discrimination)		Threshold (item difficulty)	
Item	α_i	p value	β_i	p value
LR1	0.861	0.000	−0.761	0.000
LR2	0.524	0.000	0.278	0.010
LR3	0.534	0.000	−0.671	0.000
LR4	0.232	0.001	2.216	0.002
LR5	0.503	0.000	0.084	*0.466*
LR6	0.064	*0.218*	*12.441*	*0.216*
LR7	0.348	0.001	3.727	0.001
LR9	0.658	0.000	−1.068	0.000
LR11	0.830	0.000	−0.227	0.028
LR12	0.959	0.000	−0.237	0.001
LR13	0.485	0.000	−0.403	0.000
LR14	0.555	0.000	0.571	0.000
LR16	1.223	0.000	0.181	0.048
LR17	0.919	0.000	−1.469	0.000
LR20	0.468	0.000	1.287	0.000
LR21	0.087	*0.126*	*4.422*	*0.157*

Table 20.4 CFA model fit finance (Model 2)

Model	χ^2 (df)	p	χ^2/df	RMSEA [C.I.]	CFI	TLI
1-factor finance (16 items)	136.395 (104)	0.018	1.31	0.017 [0.007–0.025]	0.919	0.907

These two items were eliminated in the next step to evaluate whether the model fits the data better. As shown in Table 20.2, the CFI and TLI can be hereby improved; the reduction of the χ^2-value is however not significant and the RMSEA is worse than in Model 1.1. Therefore, a clear decision only based on the empirical findings is hardly possible here. For further analysis, an individual decision for each modeling purpose must be made. In this case, we decide that both items are still relevant for the theoretical construct definition and kept them for further modeling.

For finance, the CFA model with all 16 items used in the test shows an acceptable global fit (see Table 20.4).

However, CFI and TLI are below the required cutoff value of 0.95, which demands for a local model fit evaluation (see Table 20.5). In this model, all factor loadings are

Table 20.5 Item parameters for Model 2 (finance)

	Factor loading (item discrimination)		Threshold (item difficulty)	
Item	α_i	p value	β_i	p value
LF1	0.444	0.000	–0.029	*0.816*
LF3	0.820	0.000	−0.551	0.000
LF4	1.017	0.000	0.513	0.000
LF8	0.393	0.000	−0.316	0.041
LF10	0.655	0.000	0.266	0.023
LF11	0.336	0.000	0.494	0.000
LF12	0.326	0.000	0.812	0.000
LF13	0.704	0.000	−0.687	0.000
LF14	0.397	0.000	−0.286	0.115
LF15	0.527	0.000	−0.276	0.032
LF17	0.643	0.000	1.377	0.000
LF18	0.642	0.000	0.763	0.000
LF19	0.324	0.000	0.057	*0.785*
LF20	0.611	0.000	−1.046	0.000
LF21	0.291	0.000	1.869	0.001
LF23	0.302	0.000	1.368	0.000

larger than zero and are all significant. We found only two items with non-significant thresholds with p values larger than 0.7 (LF1 and LF19). Further investigation of the actual thresholds of these two items showed that both parameters are close to zero. Since the item thresholds can also be interpreted as item difficulty, a value close to zero reflects an item which is neither difficult nor easy.

Therefore, for finance, the model including all items was accepted and used for further analyses.

Before we could start to test the hypothesis, it was necessary to test another model beforehand. In this next step, we tested whether accounting and finance are empirically separable areas, as theoretically characterized in the construct definition.

20.4.2.2 Two-Dimensional CFA Models

To investigate whether the data actually reflects two empirically separable factors, we tested which model fits the data better. To this end, we bundled all items from accounting and finance together to one factor (see Model 3.2 in Table 20.6) and tested that factor against a model consisting of two separate factors, one for each area (see Model 3.1. in Table 20.6). For both models, a χ^2-statistic was generated and tested for significance.

As shown in Table 20.6, as expected, the two-dimensional structure (Model 3.1) fits the data better than a one-dimensional model (Model 3.2). For Model 3.1, the RMSEA as well as the CFI and TLI statistics show better values. Moreover, despite the fact that both models show acceptable fit criteria, the χ^2-difference test revealed a significant value which indicates that the two-dimensional model (Model 3.1) fits the data better.

The theoretical assumption that the content areas of finance and accounting are empirically separable holds, and we can conclude that the WiwiKom test reliably measures the two content areas. Therefore, Model 3.1 was the baseline model for further analyses.

Table 20.6 Dimensionality of accounting and finance

Model	χ^2 (df)	*p*	χ^2/df	RMSEA [C.I.]	CFI	TLI
3.1: 2-factor accounting (14 items) and finance (16 items)	446.510 (404)	0.071	1.105	0.008 [0.000–0.013]	0.940	0.935
3.2: 1-factor accounting (14 items) and finance (16 items)	451.141 (405)	0.056	1.113	0.009 [0.000–0.013]	0.935	0.930
χ^2-difference:	5.414	0.020				

20.4.3 Two-Dimensional Multilevel Model with Covariates (MMIMIC Models)

To test the hypotheses H1 to H3, model 3.1 was extended and in the final analysis step a model with all explanatory variables was calculated as a regression model. Additionally, this model was estimated with variance separation on two levels, where the variance on institutional level (‘between’) was separated from the individual level variance (‘within’). Specifically, in Mplus (Version 8; Muthén & Muthén, 1998–2017) several multilevel multiple indicator multiple causes (MMIMIC) models were calculated (Schmidt et al., 2016). This modeling variant allows, in addition to the multilevel structure, to include the test values as a latent ability as well as to consider the answers from both content-related but nevertheless empirically separable subareas of business as two dependent variables in a common model. Based on the group differences identified, the model simultaneously takes into account the personal factors (gender, type of university attended, migration background, commercial training, school-leaving grade, and Bachelor grades). To check convergent validity with conceptually related constructs from a nomological network, as required by the AERA et al. standards (2014) for the validation process, the influences of attendance and grades of the relevant courses on the results of the respective test areas were examined.

20.4.3.1 Null Model

In a first step, as this is a multilevel analysis, the so-called null model was calculated to check how much of the total variance falls on the university level (this is expressed by the intraclass correlation coefficient [ICC]) (Hox et al., 2017). Before we start analyzing or interpreting the model parameters, we investigate some descriptive properties of our multilevel data structure. In total, there are 40 clusters according to the 27 universities and 13 universities of applied sciences to which the 1523 master’s students of economics belong (see Sect. 20.2).

As shown in Table 20.7, for most universities between 10 and 60 students were assessed. There are a few outliers with more than 100 students and an average cluster size of 37.3. This finding shows how important it is to consider the multilevel structure of our data and to take the different cluster sizes into account when testing the hypothesis. This was done by using the institution code as cluster variable and the two-level analysis option of Mplus (Muthén & Muthén, 1998–2017). This then allows us to also integrate covariates on the ‘within’ and the ‘between’ level in further modeling steps. The ‘within’ level hereby represents the characteristics for each institution whereas the ‘between’ levels represent the characteristics that are equal between all institutions. Before covariates can be interpreted, the null model had to be evaluated first.

Table 20.7 Distribution of students over the 40 universities

Cluster size (No. of students per cluster)	Institution ID
3	9
9	16
10*	17, 19
11	25, 28, 32
12	2
13	11, 6, 23
14	15, 7
15	8, 12
16	39
18	22
19	10
20	34
21	29
23	27
24	18
26	35
28	4
31	13
33	33
36	40
41	30
44	20
45	31
53	3
54	37
58	24
59	36
60	21
121	38
133	5
141	14
155	1
Average cluster size	37.3

* Some cluster sizes occur at multiple instutions

Table 20.8 Intraclass correlation of each item in the null model

Intraclass variable	Intraclass correlation	Intraclass variable	Intraclass correlation	Intraclass variable	Intraclass correlation
LF1	0.001	*LF3*	*0*	LF4	0.027
LF8	0.018	LF10	0.036	*LF11*	*0*
LF12	*0*	LF13	0.008	LF14	0.024
LF15	0.008	LF17	0.085	LF18	0.011
LF19	0.017	*LF20*	*0*	LF21	0.029
LF23	0.001	LR1	0.028	*LR2*	*0*
LR3	0.06	LR4	0.035	*LR5*	*0*
LR6	*0*	LR7	0.189	LR9	0.051
LR11	0.049	LR12	0.007	*LR13*	*0*
LR14	0.012	LR16	0.081	LR17	0.071
LR20	0.023	LR21	0.022		

Using the two-level *basic* option of Mplus (Muthén & Muthén, 1998–2017), the ICC reveals eight items with no intraclass correlation, i.e., there was no variance on the ‘between’ level (see Table 20.8).

For these eight items, the only heterogeneity was found on the individual level, and we cannot assume that these items perform differently between different universities. Therefore, we eliminated them from the ‘between’ level in further modeling steps.

20.4.3.2 MMIMIC-Models for Hypothesis Testing

After the null model showed that there was a variance greater than zero among most of the items, we started estimating the MMIMIC models to test the hypothesis H1 to H3. We followed a step-by-step approach to test the model fit for adding each explanatory variable separately. The testing of the hypotheses however can only be interpreted in a joint model that includes all covariates.

First, we evaluated the global model fit (see top of Fig. 20.1). The χ^2 statistic was not significant and therefore represented a good fit. However, CFI, TLI, and RMSEA were beyond or close to the range of the cutoff criteria: The TLI was larger than one and thus larger than the range of CFI which lies between zero and one. For TLI, Muthén and Muthén (2017) recommend to truncate if the value should be beyond one. The reason is most likely that the sample size is too small for the complex model we estimated here. However, since all criteria do not indicate a worse fit, we can assume that the model fits the data well and all model parameters can be validly interpreted.

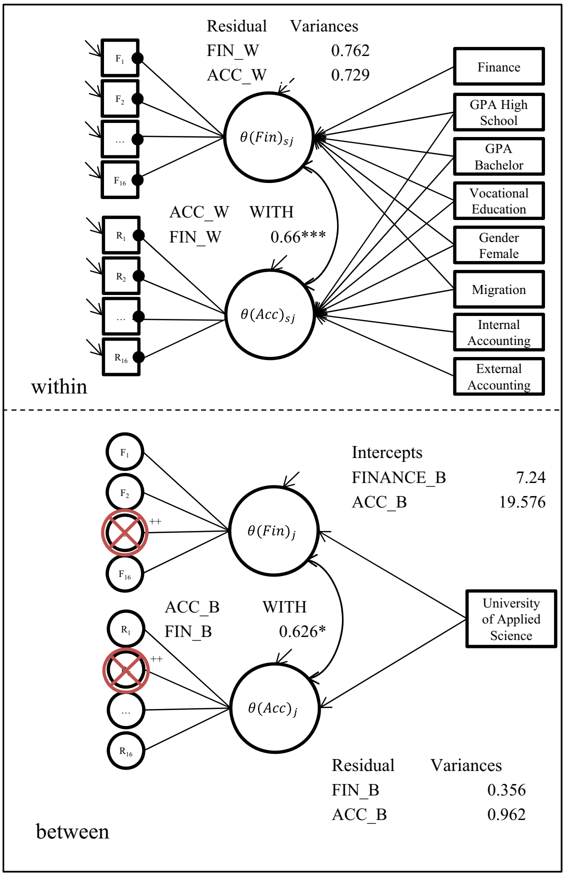
When evaluating the local model fit (see STDYX Factor Loadings on the left side of Fig. 20.1), all factor loadings were significant, except for LR6 and LR21. However, these items were not significant in the single model M1.1 (see Sect. 20.4.2) and

Chi-Square Test of Model Fit

Value	1198.406*
Degrees of Freedom	1239
P-Value	0.7914
CFI	1.000
TLI	1.195
RMSEA	0.000

STDYX Factor Loadings Within Level^{+}

LF1	0.352***
LF3	0.651***
LF4	0.656***
LF8	0.287***
LF10	0.529***
LF11	0.32***
LF12	0.253***
LF13	0.533***
LF14	0.447***
LF15	0.487***
LF17	0.494***
LF18	0.512***
LF19	0.314***
LF20	0.549***
LF21	0.32***
LF23	0.292***
LR1	0.542***
LR2	0.37***
LR3	0.33***
LR4	0.309***
LR5	0.224***
LR6	-0.007
LR7	0.285***
LR9	0.605***
LR11	0.636***
LR12	0.689***
LR13	0.426***
LR14	0.465***
LR16	0.79***
LR17	0.759***
LR20	0.326***
LR21	0.026



STDYX Regression Results

Within Level

FIN_W	ON
GPA_HS	-0.141**
GPA_BA	-0.22***
VE	0.108*
FEMALE	-0.349***
MIGRAT	-0.145***
FIN_C	0.093
ACC_W	ON
GPA_HS	-0.145***
GPA_BA	-0.215***
VE	0.05
FEMALE	-0.356***
MIGRAT	-0.188***
IACC_C	0.109**
EACC_C	0.099*

Between Level

FIN_B	ON
UAS	-0.803***
ACC_B	ON
UAS	-0.194

***p<0.01; **p<0.05; *p<0.1

$^{+}$ At Between Level Mplus standard modeling options used (according to Christ & Schlüter, 2012)

$^{++}$ At Between Level 8 Items removed from model, because ICC=0 (LF3, LF11, LF12, LF20, LR2, LR5, LR6, LR13)

Fig. 20.1 MMIMIC model with all explanatory variables

therefore this finding can be ignored. Both items are relevant in theory and therefore remained included in the model for empirical stability. However, both items were close to zero, indicating that they do not contribute to the individual level of finance and accounting knowledge measures. Hence, the relations between the predictor variables and the knowledge factors can still be interpreted.

Before we could start interpreting the effect of the predictor variables, we first considered the latent factors (see numbers around the circles on the 'within' and 'between' level in Fig. 20.1). The two factors finance and accounting were positively correlated on both levels, i.e., having a higher knowledge in one content subarea was also related to a higher knowledge in the other content subarea. This is in line with the construct definition of the WiwiKom test that both domains are subareas of business knowledge. The correlation on the 'within' level (ACC_W WITH FIN_W) was with 0.66 and a *p* value smaller than 0.01 and on the 'between' level (ACC_B WITH FIN_B) with 0.63 and a *p* value smaller than 0.1. This indicates that the relation on the 'within' level is a bit stronger than on the 'between' level, which is a plausible finding. The knowledge factors were more correlated within each university than between universities. Additionally, the results of the factor variables showed that the intercept of accounting (ACC_B) was with 19.58 much higher than for finance (FINANCE_B) with 7.24, which indicates that the general knowledge level of accounting is much higher in the whole population.

When it comes to interpreting the covariates (see the right side with STDYX regression results in Fig. 20.1), the school-leaving grade (Grand Point Average for High School = GPA_HS) and the bachelor grade (GPA for Bachelor = GPA_BA) were included in the model as control variables, since grades are generally one of the best predictors for academic performance (Hell et al., 2008). This was also the case for the knowledge of accounting and finance. Both variables have a negative influence on both factors, since the better the grade, the lower the number of the factor; i.e., having a lower number in one of the grades is associated with a higher level in both knowledge subareas. This again, reflects the importance to keep both variables as control variables in the model.

For testing hypothesis 1a, we included gender as predictor variable into the model. Female students have a lower subject-specific knowledge in both, accounting and finance, in comparison to male students. In both subareas, the FEMALE coefficient was negative and significant, indicating that H1a cannot be rejected.

The coefficient MIGRAT was also negative and significant for both, accounting and finance. Therefore, similarly, hypothesis H1b cannot be rejected, since students with a migration background have a lower level of subject-specific knowledge in accounting and finance than students without a migration background.

The findings regarding the attended courses were less straightforward: In finance, the attendance of at least one finance course did not have a significant effect on the level of finance knowledge (coefficient FIN_C was 0.093 and not significant). However, attending at least one course in internal accounting (IACC_C) was significantly related to higher levels of knowledge in accounting. The attendance of at least one course in external accounting (EACC_C) might also have a significant positive impact on knowledge in accounting, but the *p* value was only smaller than 0.1. Therefore, hypothesis 2 had to be rejected, at least for the domain of finance.

Hypothesis 3 stated that prior knowledge should be positively associated with knowledge in the master's studies in business and economics. The findings showed that vocational education in the field of business only had a slight impact on the knowledge in finance and no effect on the knowledge of accounting. Therefore, H3 must also be rejected, at least for accounting.

20.5 Discussion and Conclusion

In line with expectations, the test performance correlated both with previous attendance of relevant subject-specific lectures and with personal influencing factors, gender, bachelor's degree, and school-leaving grades consistently showing significant effects in all analyses.

Male students performed significantly better than female students. This supports H1a and leads to the question of whether female students generally perform worse in business and economics due to factors such as a multiple-choice task format or economics test interest, which were often discussed in prior research (e.g., Brückner et al., 2015a, 2015b). For instance, Brückner et al. (2015a) found that economic test questions without numerical content resulted in less differences in the performance of male and female students than questions with numerical content.

Students with parents with migration background and/or a family language other than German had a negative effect on the knowledge test results (H1b), which is in line with prior research, according to which foreign language learners or migrants achieve worse results in economic knowledge questions (e.g., Förster et al., 2015; Happ et al., 2019). This raises the question of whether the worse test performance is a result of systematic test bias among foreign language students in the processing of tasks and/or the acquisition of knowledge during their university studies (for a critical discussion, see Schlax et al., 2020). For instance, poorer (test) language comprehension might make it more difficult to complete the test items and the students would possibly be able to achieve results comparable to those of their German counterparts if, for example, they were to complete the test over a longer period of time. In addition, migrant students may have greater difficulty in acquiring the subject-specific knowledge during their university studies due to linguistic and cultural differences, which could lead to comparatively lower economic knowledge (Happ et al., 2019). In this respect, further investigations in a systematic manner through appropriate operationalization, such as more time for processing or language-free tasks are required.

In addition to the control of the personal influencing factors, the correlations found between the attended courses and knowledge test performance were in line with expectations for the subarea for accounting, but not for finance. Therefore, the hypothesis that attending at least one course as a significant indicator for the prediction of the acquisition of subject-specific knowledge must be rejected for the subarea of finance (H2). However, the level of knowledge in the area of accounting showed the assumed correlation with the attendance of the corresponding lectures. This suggests that attending at least one course related to this subject (internal or external accounting) leads to a significant increase in corresponding knowledge, which is reflected in a better test result. Thus, at least for one content area, the results confirm the relationship of the assessed knowledge with lectures attended as conceptually related constructs. With regard to this (validation) criterion (AERA et al., 2014), it can therefore be assumed that the procedure developed and used within the framework of WiwiKom is also suitable for application to students at the beginning of the Master's program.

In this context, while prior vocational education only had a minor effect on knowledge of finance and no effect on knowledge of accounting, a greater correlation between economic knowledge in both domains and the bachelor's grade than with the school-leaving grade became evident. Therefore, the influence of the bachelor's grade, consisting of the module grades of various economic subjects was, as expected, consistently greater than that of the school-leaving grade, which mainly covers achievements in non-economic subjects (Happ et al., 2019). Thus, the greater influence from subject-relevant academic education compared to the influence from school education and vocational training became evident here.

Overall, the results suggest that the WiwiKom test is suitable for the diagnosis of economic knowledge throughout the course of the master's study. Thus, the test represents a method for the valid entrance and process diagnostics of the knowledge development of master students, which is urgently required, especially with regard to the increasing heterogeneity of the student body and the specific study entrance requirements. It is an important step toward meeting the need for reliable assessments, identifying the strengths and demands of master's students, and challenging them to apply their knowledge in practice-oriented situations.

With regard to the MMIMIC approach, however, and in addition to the demonstrated advantages when using this method for educational assessment purposes, there are some limitations. The clear advantage of this approach is the integration of all personal, contextual, and item-related information into one model, which goes in hand with no information loss. If a latent path model is modeled using already aggregated scores, such as sum scores for the knowledge scales, there is a potential loss of information regarding item difficulty and discrimination for each student observed in the sample. However, latent models are generally complex in terms of the numbers of parameters that have to be estimated. This requires big sample sizes. When it comes to multilevel latent models, it needs to be ensured that the models used on all levels are well-fitting and adequate. Otherwise, there is a risk that no model convergence could be reached, and no parameters could be estimated at all. In our case, the TLI for model fit had to be truncated due to a too high value greater than one. This could be a sign that the sample size is too small, and/or the model too complex. However, despite this fact, all model parameters could still be estimated and on item level significant results were generated. Additionally, since iterative algorithms are used for model estimation, it can take a really long time to find an optimal solution and generate the model parameters. This was also evident in our case.

Despite these and other limitations, our study demonstrates that a multilevel approach and latent analyses present a suitable approach to explain and significantly predict the master's students' domain-specific knowledge in accounting and finance at least partly from the attendance of relevant courses, while controlling for other, theoretically expected influencing variables (such as gender or socio-economic background) in the multilevel analyses.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* American Psychological Association.

Bellin, N. (2009). *Klassenkomposition, Migrationshintergrund und Leistung – Mehrebenenanalysen zum Sprach- und Leseverständnis von Grundschülern* [Class composition, migration background and performance—Multilevel analyses on primary school students' reading comprehension]. Verlag für Sozialwissenschaften.

Borde, S. F. (2017). Student characteristics and performance in intermediate corporate finance. *Journal of Financial Education, 43*(1), 1–13.

Brahmasrene, T., & Whitten, D. (2001). Assessing success on the uniform CPA exam: A logit approach. *Journal of Education for Business, 77*(1), 45–50.

Broecke, S., & Nicholls, T. (2007). *Ethnicity and degree attainment (Research Report No. RW92).* Department of Education and Skills.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research.* Guilford Publications.

Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., Happ, R., Walstad, W. B., Yamaoka, M., & Asano, T. (2015a). Gender effects in assessment of economic knowledge and understanding—Differences among undergraduate business and economics students in Germany, Japan, and the United States. *Peabody Journal of Education, 90*(4), 503–518.

Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., & Walstad, W. B. (2015b). Effects of prior economic education, native language, and gender on economic knowledge of first-year students in higher education: A comparative study between Germany and the USA. *Studies in Higher Education, 40*(3), 437–453.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* Routledge Academic.

Byrne, M., & Flood, B. (2008). Examining the relationships among background variables and academic performance of first year accounting students at an Irish University. *Journal of Accounting Education, 26*(4), 202–212. https://doi.org/10.1016/j.jaccedu.2009.02.001

Centro Nacional de Evaluación para la Educación Superior (CENEVAL). (2010). *Guía para el sustentante. Examen General para el Egreso de la Licenciatura en Administración. EGEL-ADMON* (2nd ed.). Centro Nacional de Evaluación para la Educación Superior (CENEVAL).

Christ, O., & Schlüter, E. (2012). Strukturgleichungsmodelle mit Mplus. In Strukturgleichungsmodelle mit Mplus. Oldenbourg Wissenschaftsverlag.

Dochy, F. J. R. C., De Ridjt, C., & Dyck, W. (2002). Cognitive prerequisites and learning: How far have we progressed since bloom? Implications for educational practice and teaching. *Active Learning in Higher Education, 3*(3), 265–284.

Duff, A. (2004). Understanding academic performance and progression of first-year accounting and business economics undergraduates: The role of approaches to learning and prior academic achievement. *Accounting Education, 13*(4), 409–430. https://doi.org/10.1080/0963928042000306800

Federal Institute for Vocational Education and Training (ed.). (2018). *VET Data Report Germany 2016/2017. Facts and analyses to accompany the Federal Government Report on Vocational Education and Training—Selected findings.* VET. https://www.bibb.de/veroeffentlichungen/en/publication/show/9550

Förster, M., Brückner, S., & Zlatkin-Troitschanskaia, O. (2015). Assessing the financial knowledge of university students in Germany. *Empirical Research in Vocational Education and Training, 7*(6), 1–20.

Fritsch, S., Berger, S., Seifried, J., Bouley, F., Wuttke, E., Schnick-Vollmer, K., & Schmitz, B. (2015). The impact of university teacher training on prospective teachers' CK and PCK—A comparison between Austria and Germany. *Empirical Research in Vocational Education and Training, 7*(1), 133. https://doi.org/10.1186/s40461-015-0014-8

Gracia, L., & Jenkins, E. (2010). A quantitative exploration of student performance on an undergraduate accounting programme of study. *Accounting Education, 12*(1), 15–32.

Happ, R., Nagel, M., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2019). How migration background affects master degree students' knowledge of business and economics. *Studies in Higher Education,* 1–16 (online first). https://doi.org/10.1080/03075079.2019.1640670

Hell, B., Trapmann, S., & Schuler, H. (2008). Synopse der Hohenheimer Metaanalysen zur Prognostizierbarkeit des Studienerfolgs und Implikationen für die Auswahl- und Beratungspraxis. In H. Schuler & B. Hell (Eds.), *Studierendenauswahl und Studienentscheidung* (pp. 43–54). Hogrefe.

Helmke, A., & Schrader, F.-W. (2001). School achievement, cognitive and motivational determinants. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 20, pp. 13552–13556). Pergamon.

Helmke, A., & Schrader, F.-W. (2011). Vom Angebots-Nutzungs-Modell zur Unterrichtsentwicklung. In A. Bartz, H.-J. Brandes, & S. Engelke (Eds.), *Praxishilfen für die mittlere Führungsebene in der Schule* (pp. 3–6). Carl Link Verlag.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications.* Routledge.

Keef, S. P., & Roush, M. L. (1997). New Zealand evidence on the performance of accounting students: Race, gender and self-concept. *Issues in Accounting Education, 12*(2), 315–330.

Kim, H. H., & Lalancette, D. (Eds.). (2013). *Literature review on the value-added measurement in higher education.* OECD.

Kraitzek, A., Förster, M., & Zlatkin-Troitschanskaia, O. (2020). Influences on Master's Degree Students' Economic Knowledge. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, & C. Lautenbach (Eds.), *Student learning in German Higher Education* (pp. 401–429). Springer VS.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2017). *Unusual TLI value*s. Accessed via https://www.statmodel.com/download/TLI.pdf

Oberrauch, L., & Kaiser, T. (2019). Economic competence in early secondary school: Evidence from a large-scale assessment in Germany. *International Review of Economics Education, 35.* https://doi.org/10.1016/j.iree.2019.100172

OECD. (2017). *Education at a glance 2017: OECD indicators.* OECD Publishing.

OECD. (2020). *Education at a glance 2020: OECD indicators.* OECD Publishing.

Owen, A. L. (2012). Student characteristics, behavior, and performance in economics classes. In G. M. Hoyt & K. McGoldrick (Eds.), *International handbook on teaching and learning economics* (pp. 341–350). Edward Elgar.

Park, L. J., & Hayes, R. S. (1994). Men and women: Equal in accounting? *Journal of Education for Business, 69*(6), 349–353. https://doi.org/10.1080/08832323.1994.10117712

Paver, B., & Gammie, E. (2005). Constructed gender, approach to learning and academic performance. *Accounting Education, 14*(4), 427–444. https://doi.org/10.1080/06939280500347142

Schlax, J., Zlatkin-Troitschanskaia, O., Kühling-Thees, C., & Brückner, S. (2020). Influences on the development of economic knowledge over the first academic year. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, & C. Lautenbach (Eds.), *Student learning in German Higher Education.* Springer VS. https://doi.org/10.1007/978-3-658-27886-1_19

Schmidt, S., Zlatkin-Troitschanskaia, O., & Fox, J.-P. (2016). Pretest-posttest-posttest multilevel IRT modeling of competence growth of students in higher education in Germany. *Journal of Educational Measurement, 53*(3), 332–351. https://doi.org/10.1111/jedm.12115

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Solga, H., Protsch, P., Ebene, C., & Brzinsky-Fay, C. (2014). *The German vocational education and training system: Its institutional configuration, strengths, and challenges.* Paper presented at the Wissenschaftszentrum Berlin für Sozialforschung [Berlin Social Science Center], Berlin.

Statistisches Bundesamt. (2018). *Anzahl der Studierenden an deutschen Hochschulen in den 20 am stärksten besetzten Studienfächern im Wintersemester 2017/2018* [Graph]. In Statista. Access at October 07, 2019, from https://de.statista.com/statistik/daten/studie/2140/umfrage/anzahl-der-deutschen-studenten-nach-studienfach/

Terry, A. (2002). Student performance in the introductory corporate finance course. *Journal of Financial Education, 28*, 28–41.

Tse, H., & Tam, K. L. (2017). Getting the basics right: Factors shaping student performance in intermediate economics. *Economic Analysis and Policy, 53*, 1–8.

Weber, B. (2002). Economic Education in Germany. *Journal of Social Science Education, 2.* https://doi.org/10.4119/jsse-267

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from the German assessment of business and economics competence. In H. Coates (Ed.), *Assessing learning outcomes: Perspectives for quality improvement* (pp. 175–197). Lang.

Zlatkin-Troitschanskaia, O., Förster, M., Schmidt, S., Brückner, S., & Beck, K. (2015). Erwerb wirtschaftswissenschaftlicher Fachkompetenz im Studium. Eine mehrebenenanalytische Betrachtung von hochschulischen und individuellen Einflussfaktoren. In S. Blömeke, & O. Zlatkin-Troitschanskaia (Eds.), *Kompetenzen von Studierenden* (pp. 116–134). Beltz.

Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., Förster, M., & Brückner, S. (2019). Validating a test for measuring knowledge and understanding of economics among university students. *Zeitschrift Für Pädagogische Psychologie, 33*(2), 119–133.

Zlatkin-Troitschanskaia, O., Schmidt, S., Brückner, S., Förster, M., Yamaoka, M., & Asano, T. (2016). Macroeconomic knowledge of higher education XE “higher education (HE)” students in Germany and Japan—A multilevel analysis of contextual and personal effects. *Assessment & Evaluation in Higher Education, 41*(5), 787–801.

Zorlu, A. (2011). *Ethnic disparities in degree performance (Discussion Paper No. 6258).* IZA.

Dr. Susanne Schmidt is a Research Associate at the Chair of Business and Economics Education at Johannes Gutenberg University (Germany), where she earned her doctoral degree in 2016. Her research focuses and expertise lies in analyzing longitudinal studies, including multilevel and process data, to describe the learning processes of higher education students.

Professor Olga Zlatkin-Troitschanskaia has been Chair of Business and Economics Education at Johannes Gutenberg University (Germany) since 2006. She has directed numerous research projects on modeling and measuring student knowledge, skill development, and learning outcomes in higher education at both the national and international levels.

Marie-Theres Nagel is a Research Assistant and the Chair of Business and Economics Education at Johannes Gutenberg University (Germany), where she is currently working on her Ph.D. thesis. In 2020 she became a junior member of the Gutenberg-Academy, which funds and fosters the most outstanding Ph.D. candidates from all departments of Johannes Gutenberg University.

Index

A

Academic performance, 6, 7, 240, 256–260, 263, 265–268, 364, 366, 369, 370, 378, 418

Achievement, 6, 30, 51, 55, 57, 59, 62, 71, 72, 74, 76, 78, 80, 82, 84, 85, 87, 88, 91, 94, 101, 102, 108, 123–127, 129, 131–133, 142, 145, 146, 162, 163, 165, 186–194, 196, 197, 199, 211, 212, 215, 218, 220, 223, 225, 229, 236–239, 244, 246–250, 257, 262, 263, 271–273, 277–279, 284, 285, 287, 288, 292, 298, 300–304, 368, 420

Adolescents, 256, 257, 263, 267, 268

Akaike Information Criterion (AIC), 20, 24, 25, 30, 34, 37–39, 169, 172, 296, 297, 345, 351, 352, 355, 357, 375–377

B

Bayesian Information Criterion (BIC), 20, 24, 25, 30, 34, 37–39, 169, 172, 296, 297, 345, 351, 352, 355, 357

C

Climate *vs.* contextual effects, 114

Composition effects, 342, 345, 358

Country contexts, 318, 319, 330–333

Cross-level effect, 6, 196, 321, 330, 331

Cross-national analysis, 143

Cultural status, 186, 256, 258, 267

D

Data visualization, 5, 71, 72, 74, 76, 78, 80, 82, 94

Design-based approach, 16

Design effects, 15–17, 21, 23–25, 144

Direct instruction (DI), 6, 237–239, 241, 244, 246–250

Domain-specific knowledge, 8, 404–407

Doubly latent multilevel structural equation modeling (DL-MLSEM), 5, 101, 107, 114

E

Effective schools, 7, 272, 273, 275–279

Estimation, 4, 12, 17, 19–21, 25, 30–32, 34, 36, 37, 51, 55, 103, 104, 123, 124, 127, 129, 144, 145, 165, 168, 171–174, 176, 177, 191, 194, 195, 199, 242, 255, 256, 258, 373, 387, 421

G

Gender, 7, 58, 62–64, 86, 102, 107, 126, 127, 129, 131, 163, 191, 193, 208, 211, 241, 243, 245, 246, 256, 258, 259, 263, 265, 267, 288, 316, 326, 364, 370–374, 376–378, 388–390, 403–405, 408, 409, 414, 418, 419, 421

Globalization, 284

Graph Builder, 76, 91, 93

M. S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research*,
https://doi.org/10.1007/978-981-16-9142-3

H

Hierarchical data, 3, 71, 95, 100, 242, 370

Hierarchical Linear Modeling (HLM), 5, 12, 36, 50, 54, 58, 65, 70, 100, 174, 239, 322, 323

Higher education (HE), 7, 207, 325, 329, 364, 366, 367, 371, 372, 378, 407, 410

High School and Beyond (HSB), 5, 71, 85

HLM model, 330

Home background factors, 273–275

Hypotheses, 34, 36, 106, 110, 114, 143, 162, 186, 191, 214, 218, 240, 250, 285, 314, 406, 414, 416

I

Inquiry-based teaching (IBT), 6, 236–241, 246–250

Interclass coefficient, 187

International large-scale assessment (ILSAs), 5, 121, 141–149, 154, 155, 186, 197, 271

International large-scale studies, 5

International Standard for Educational and Psychological testing, 11, 408

K

Kindergartens, 7, 8, 163, 340, 384–393, 395–398

L

Language skills, 7, 260, 341–343, 345, 350, 355, 358

Large-scale assessment(s), 122, 123, 135, 155, 199

Longitudinal study, 71, 162

M

Mathematics achievement, 6, 7, 72, 212, 213, 217, 223, 225, 228, 229, 271–279

Math teachers, 62

Measurement and sampling errors, 105, 114, 193

Measurement model, 55, 101, 103, 104, 110, 111, 113, 293, 297, 298

Model-based procedure, 16, 17

Model fit, 15, 16, 20–22, 24, 25, 29, 30, 33, 35, 38–40, 45, 53, 54, 63, 73, 86, 100, 104, 109–111, 114, 167, 291, 299, 345, 352, 410, 412, 413, 416, 420

Model selection, 4, 29, 30, 33, 37, 39, 170

Model specification, 5, 6, 124–126, 154, 162, 165, 169, 176, 178, 353

Multi group confirmatory analysis, 293

Multilevel analysis, 4–7, 14, 65, 88, 147–150, 153–155, 187, 191, 213, 218, 223, 240, 274, 276, 313, 331, 345, 351, 414

Multilevel confirmatory factor analysis (MCFA), 4, 12, 13, 15–26, 103, 104, 110, 167

Multilevel data, 13–15, 37, 50, 64, 65, 100, 106, 107, 161, 164, 165, 172, 414

Multilevel mediation, 106, 107

Multilevel modeling (MLM), 3–8, 12, 14, 30–32, 36, 39, 40, 42–44, 50, 55, 65, 70, 71, 73, 78, 84, 94, 100, 105, 106, 114, 122, 125, 129, 135, 161–163, 165, 172–177, 185, 186, 189, 190, 193, 194, 199, 218, 241, 272, 330, 343–345, 351

Multilevel research, 106, 162, 165–167, 312

Multilevel Structural Equation Modeling (MLSEM), 4, 50, 51, 53–55, 65, 100, 105, 107

Multivariate multilevel analysis, 6, 58, 62

N

Nested data, 14, 15, 17, 50, 51, 65, 71, 85, 94, 95, 105, 161, 211, 242, 409, 410

Non-native children, 341, 343, 345, 353, 355, 358

P

Plausible values, 5, 6, 123, 124, 127, 129–133, 135, 144–147, 154, 155, 168, 186, 187, 190–193, 199, 218, 258, 259, 274–276, 288, 299, 300

Professional development (PD), 53, 54, 154, 207–210, 216–218, 222–225, 228, 229, 246, 249, 288, 292, 298, 303

Programme for International Student Assessment (PISA), 4–6, 121–125, 131, 132, 135, 141–143, 145–147, 149, 151, 155, 186–190, 192, 194–199, 235–237, 239–242, 244, 247–250, 255, 256, 258–260, 262, 267, 268, 321

Progress in International Reading Literacy Study (PIRLS), 142, 143, 145–148, 150, 151, 153, 155, 190

R

R, 5, 36, 45, 123, 125–127, 129, 130, 132, 133, 135, 242

Replicate weight, 6, 124, 154, 186, 197–199, 218

Reporting guidelines, 55

Reporting practice, 6, 162

R package, 5, 166, 172, 173, 175, 176

S

Satisfaction survey, 7

School ethical culture (SEC), 7, 284, 285, 287–289, 291, 296, 299–301, 303, 304

School level factors, 273, 277, 279

Science, Technology, Engineering, and Mathematics (STEM), 7, 364, 366, 367, 369, 370, 372, 373, 375, 377–379

Scientific literacy, 6, 237, 239, 240, 242, 246, 248, 249

Segregation, 194, 206, 207, 340–342, 358, 370

Service quality, 8, 384, 387, 390, 397, 398

Socio-economic status, 147, 256, 258, 260, 263, 267

Subject-specific knowledge, 404–408, 418–420

Survey weight, 6, 17, 123, 193–196, 199

T

Teacher qualification, 206, 207, 211–215, 218, 220, 223, 225, 228, 229

Teachers, 5–7, 13, 22, 23, 51, 53, 54, 70, 101, 102, 107–111, 113, 143, 147, 148, 150, 162, 164, 186, 190, 199, 206–218, 220–223, 225, 226, 228, 229, 236, 237, 239–241, 246, 248–250, 260, 272, 279, 283–289, 291–304, 311–323, 325, 327–333, 347, 368, 386

Teacher turnover intention, 312, 314, 315, 317, 318, 320–322, 325, 328, 330–333

Three-level model, 6, 55, 56, 58, 147, 151, 153, 190, 196, 199, 376

Trends, 4, 6, 64, 74, 94, 210, 218, 229, 271, 317, 369, 393

Trends in International Mathematics and Science Study (TIMSS), 4–7, 52, 142, 143, 145–148, 150, 151, 153, 155, 186, 190, 196, 197, 210, 212–216, 220–225, 229, 238, 239, 271, 273–279, 284, 285, 287–289, 299, 303, 304

V

Variance decomposition, 39, 41, 187, 189, 196, 199