Title
Multilevel Factor Analysis and Student Ratings of Instructional Practice

Author
Schweig, Jonathan David

Publication Date
2014

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Multilevel Factor Analysis and Student Ratings of Instructional Practice

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Education

by

Jonathan David Schweig

2014

Abstract of the Dissertation

Multilevel Factor Analysis and Student Ratings of Instructional Practice

by

Jonathan David Schweig

Doctor of Philosophy in Education

University of California, Los Angeles, 2014

Professor José Felipe Martínez, Chair

Student surveys of classroom climate can provide teachers, administrators, and researchers with valuable information about instructional practice and are becoming a critical component in policy efforts to assess and improve teaching. Seventeen states and many large municipalities including Chicago, Illinois, Memphis, Tennessee and Denver, Colorado include student surveys as a key component in formative and summative teacher evaluation plans. Several other states and municipalities—including Los Angeles—are in the process of developing or piloting student surveys for future use in teacher evaluation. Advocates note that students are natural observers of their classroom environments, have extensive and rich knowledge of their teachers, and that student ratings can be predictive of important outcomes, such as student academic and socio-emotional development. In addition, student surveys are relatively easy and cost-effective to administer. At the same time, using information from student surveys for formative or summative assessment of teachers presents conceptual and methodological challenges. Inferences about a teachers instructional practice are often based on aggregated student survey responses, and a key step in assessing the appropriate uses of the information collected from student surveys is to understand the dimensions of classroom climate or instructional practice that are discernible when looking at

student responses aggregated by classroom.

This dissertation proposes a new approach for exploring the dimensionality of aggregated student ratings. This approach also has the potential to provide validity evidence supporting the use of student surveys as measures of instructional practice in both formative and summative evaluation. Specifically, this dissertation applies a non-parametric cluster-bootstrap technique to a multilevel factor analysis framework that allows researchers to evaluate psychometric models where data is collected from students but teachers are the object of measurement. This approach can be extended directly to applications where teachers are clustered within schools. Four research topics were investigated:

1. The efficiency of the proposed approach compared to other possible approaches to analyzing the teacher-level covariance structure.

2. The comparative performance of the proposed approach and other possible approaches, in terms of the accuracy of parameter estimates, consistency of standard errors, and distribution of test-statistics for model appraisal.

3. The extension of the proposed approach to datasets with three levels (e.g., where students are clustered in classrooms, and classrooms are clustered in schools).

4. The application of the bootstrap method to a realistic dataset to illustrate how they may be used to investigate the dimensions of teacher practice that are discernible from a student survey of instructional practice.

The first three research topics are investigated using a series of simulation studies. These studies involve a range of simulation conditions reflecting conditions commonly encountered in surveys of instructional practice. The cluster-bootstrap technique was then applied to data collected from the New Mexico Opportunity to Learn survey in order to illustrate how the technique can be used to make

inferences about the discernible dimensions of instructional practice, and about how survey-derived variables predict student achievement growth in math and reading.

The findings of this dissertation contribute both to the methodological and substantive literatures. Methodologically, the results demonstrate that the proposed cluster-bootstrap technique can be used in conjunction with maximum likelihood estimation to yield accurate parameter estimates, and that for sufficiently large sample sizes, test statistics and standard errors based on the cluster bootstrap technique will yield valid inferences about the psychometric properties of aggregated survey responses. The simulation study also demonstrated that the cluster bootstrap technique can be extended to three-level data sets where students are clustered in classrooms, and classrooms are clustered in schools.

In addition, these results offer some of the first empirical evidence of how covariance structure analysis may be applied to student surveys of instructional practice, when the clustering of teachers into schools is acknowledged. This is of particular importance for applied researchers or policy makers using aggregated student surveys for formative or summative evaluation. In this context, understanding the constructs measured by aggregated survey responses is a critical step in developing and testing theories about how dimensions of classroom practice relate to student academic and socio-emotional development, and it is critical step in building systems that can provide high-quality diagnostic feedback to teachers about their instructional practice. In the case of the New Mexico Opportunity to Learn survey, it was shown that students are able to distinguish three dimensions of instructional practice, but that only overall ratings of instructional practice were predictive of student achievement growth.

The dissertation of Jonathan David Schweig is approved.

Peter Bentler

Li Cai

Noreen Webb

José Felipe Martínez, Committee Chair

University of California, Los Angeles

2014

To Jenn and Fiona ...

for all of their love, patience and support. And to my mother and father for encouraging my curiosity. I love you all very much.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I gratefully acknowledge the guidance of José Felipe Martínez, who provided tremendous support and assistance throughout my studies and through the process of researching and writing this dissertation. I would also like to acknowledge Peter Bentler, Li Cai and Noreen Webb for their patience and willingness to answer what must have seemed at times like a relentless torrent of questions.

I am grateful to Pete Goldschmidt and the New Mexico Public Education Department for their support, and for providing access to the data used in this dissertation.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B080016 to the University of California, Los Angeles. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Finally, I am grateful to the many faculty, colleagues, and friends at UCLA that I have had the privilege of working with over the past five years. In particular, I would like to thank Meredith Phillips, Kyo Yamashiro, James Stigler, Jia Wang, Mark Hansen, Scott Monroe, Larry Thomas, Jordan Rickles, Megan Kuhfeld, Alejandra Priede, Lisa Dillman, Belinda Thompson, and Patricia Quiñones.

Vita

2014	M.S. (Statistics), University of California, Los Angeles.
2002	M.A. (Curriculum and Teacher Education), Stanford University.
1999	A.B. (Mathematics & English and American Literature), Brown University.
2009–present	Graduate Student Researcher, National Center for Research on Evaluation, Standards, and Student Testing, UCLA.
2005–2009	Program Manager and Program Director, Math for America, New York, NY.
2002–2005	Mathematics Teacher, Nightingale-Bamford School, New York, NY.
1999–2001	Mathematics Teacher, Lincoln School, Providence, RI.

Publications and Presentations

Schweig, J. (2014) Quantifying error in survey measures of school and classroom environments. *Applied Measurement in Education*. doi:10.1080/08957347.2014.880442

Schweig, J. (2013) Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*. doi:10.3102/0162373713509880

Herman, J., Wang, J., Straubhaar, R., Schweig, J. , Hsu, V. (2013) Evaluation of Green Dot's Locke transformation project: From the perspective of teachers and administrators. *CRESST report 824.*

Wang, J. , Schweig, J., Griffin, N., Baldanza, M., Rivera, N. Hsu, V.(2013) Inspiring Minds through a Professional Alliance of Community Teachers (IMPACT): Evaluation results of the Cohort I math and science apprentice teachers. *CRESST report 826.*

Downer, J., Stuhlman, M., Schweig, J., Martinez, J. F. (2013) Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. Presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Schweig, J. (2012) Multilevel Construct Validity: An empirical investigation of the assumption of cross-level invariance. Poster presented at the Modern Modeling Methods Conference, Storrs, CT.

Schweig, J. (2012) Policy implications of different approaches to describing the accuracy of school and classroom environment measures. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

CHAPTER 1

Introduction

The development of teacher evaluation systems has become one of the key challenges facing the U.S. education system (American Federation of Teachers, nd; Bill and Melinda Gates Foundation, 2010; Weisberg et al., 2009). The fact that "teacher evaluation has now scurried onto American education's center stage"(Popham, 2013, p.3) reflects concerns about student performance on national and international assessments (Feuer, 2012; National Academies, 2010) and its implications for long-term American economic growth and development, and concerns about the persistence of achievement gaps and inequities for at risk students, who are more likely to be taught by the least experienced and least qualified teachers (Weisberg et al., 2009; Akiba, LeTendre, & Scribner, 2007; Haycock, 2001). Recent attention to teacher evaluation has also been catalyzed by federal policies, including the Obama administration's Race to the Top program, enacted in 2009, and the Elementary and Secondary Education Act (ESEA) Flexibility Program (also known as the No Child Left Behind (NCLB) waivers), enacted in 2011, both of which required state and local education agencies to form—or reform—their teacher evaluation systems in order to qualify for federal funding or NCLB waivers. For all of these reasons, (Cochran-Smith, 2010) declared "at least as far as education goes, we live in an age of accountability"(p. xiii).

The notion that teachers can be held accountable for student learning or other student outcomes has gained widespread policy and cultural acceptance. Cochran-Smith noted that terms like "*outcomes, results, consequences, effectiveness, impact,*

bottom lines, what works, empirical research base, and *evidence* have been stitched so seamlessly into the logic of the discourse that they are now unremarkable" (p. xiii). This was not always the case. Darling-Hammond (1990) noted that in the past, "improving the quality of teachers [had] not been seen as critical for improving the quality of education" (p. 17). However, there is a growing research base demonstrating that student achievement varies significantly across classrooms and teachers, and that formal teacher qualifications (i.e., degrees, and credentials) do not help explain this variation (Baker et al., 2010; Rowe, 2003). Recent research has shown that teachers are the largest "within school influence on student learning" (Haertel, 2013, p. 5) and that teachers account for a meaningful portion of the variance in student achievement (e.g., Goldhaber, Brewer, & Anderson, 1999; Nye, Konstantopoulos, & Hedges, 2004). Because of the emerging research consensus that "teachers matter" (Bill and Melinda Gates Foundation, 2010, p. 1) finding methods to measure teacher effectiveness and instructional practice has become an issue of critical importance for the development and refinement of teacher evaluation systems in state and local school districts across the country.

In addition to these factors, the desire to form or reform teacher evaluation systems is motivated by the growing sense that existing teacher evaluation systems are a "perfunctory exercise" (Bill and Melinda Gates Foundation, 2010, p. 10) of little formative (or informative) value either for the teachers, or for state or local education agencies (Glazerman et al., 2010). In fact, there is a long history of dissatisfaction in the United States with teacher evaluation systems. Weisberg et al. (2009) noted that, under most evaluation systems:

> Excellent teachers cannot be recognized or rewarded, chronically low-performing teachers languish, and the wide majority of teachers performing at moderate levels do not get the differentiated support and development they need to improve as professionals (p. 6).

Darling-Hammond (1990) noted, "teacher evaluation has often had little influence on decisions about personnel, staff development or the structure of teaching"(p. 17). Haefele (1993) noted that "the dominant model of teacher evaluation is in trouble" (p. 21). In the 1980s, Medley, Coker, and Soar (1984) wrote, that teacher evaluation was "entirely inadequate" (p. 29) to identify competent teachers, and to diagnose incompetence. In the first edition of *Teacher Evaluation: A Comprehensive Guide to New Directions and Practices*, Peterson (1995) noted "teacher evaluation as practiced in the overwhelming majority of school districts in this country consists of wrong thinking and doing" (p. 3). The second edition, published in 2000, contains the exact same quote (Peterson, 2000, p. 3), possibly indicating how little had changed in the overall landscape of teacher evaluation. Blumberg (1974) noted that teacher evaluation "tends to be a ritualized, sterile process that bears little relationship to the learning of youngsters" (p. 5).

1.1 The complexity of teaching and the complexity of teacher evaluation

While it may be true that "everyone agrees that teacher evaluations are broken" (The New Teacher Project, 2010, p. 1), reaching consensus around the specific aspects of teachers' professional or instructional practice to evaluate poses an immediate conceptual challenge (e.g., Schoenfeld, 1999) in the development of a teacher evaluation system. Specifically, it is difficult to reach consensus about what defines teaching quality, and what the corresponding evaluative criteria should be. Teaching is a complex, multidimensional activity (Shulman, 1987) situated in a "relatively ill-structured, dynamic environment"(Leinhardt & Greeno, 1986, p. 75) and interdependent with temporal, social, and cultural contexts. It has been accepted that "teachers matter", but there are many different definitions of teacher quality, and little consensus about the aspects of teaching that are connected to

teacher quality (Cochran-Smith, 2010). Cochran-Smith (2010) used two examples to illustrate how differently the work of teaching can be defined. At one end of the spectrum, Cochran-Smith cited Hanushek (2002), who defined teaching in the following way: "good teachers are the ones who get large gains in student achievement for their classes; bad teachers are just the opposite" (Hanushek, 2002, p. 2–3, in Cochran-Smith, 2010, p. xv). On the other end of the spectrum is the definition by Fenstermacher and Richardson (2005):

> By good teaching, we mean that the content taught accords with disciplinary standards of adequacy and completeness, and that the methods employed are age appropriate, morally defensible, and undertaken with the intention of enhancing the learner's competence with respect to the content studied (Fenstermacher & Richardson, 2005, p. 191, in Cochran-Smith, 2010, p. xvi).

Cochran-Smith (2010) noted that these definitions span the continuum from "simple, linear and causal" to "complex, nuanced, and contingent" (p. xvi). Along with the multiplicity of definitions of quality teaching, there are also multiple evaluative criteria that can be used as factors to determine teaching quality. Popham (2013) noted that the "single, most important decision to be made as we evaluate teachers" (p. 37) was to adopt of a set of evaluative criteria. Popham (2013) noted that there can be a great deal of variation in terms of the evaluative criteria used to make determinations about teacher quality across evaluation systems, and noted that teacher quality is often related to a variety of criteria, including instructional practice, participation in professional development activities, and community relationships. Peterson (1987) also noted that teacher performance could include work teachers do outside the classroom, including relations with parents, administrators, and other teachers, school citizenship, and contributions to the community. Shulman (1987) noted that quality teaching is directly influenced by a variety of factors including curriculum content, instructional goals, and

pedagogical models on one hand, but also student characteristics and features of the classroom and school context. Finally, there is increasing awareness of the need to explicitly consider student learning, in particular whether students achieve high rates of growth, as a key factor in assessing teacher effectiveness (Federal Register, 2009). In the current policy environment, "student's test performances, as never before, are now to become a major consideration when determining a teacher's quality"(Popham, 2013, p. 8).

1.2 The case for multiple measures

Because teaching is a complex task, and because there can be multiple associated evaluative criteria of quality teaching, it seems sensible to use "multiple evidence sources rather than only one" (Popham, 2013, p. 40) in order to make evaluative decisions about a teacher's quality. Many state and local education agency plans for teacher evaluation systems specifically call for the use of multiple measures, and the Race to the Top legislation states that "effectiveness" should be defined based on input from multiple measures (Federal Register, 2009). Sound teacher evaluation systems must rely on multiple complementary indicators in order to be valid, comprehensive, and useful (Baker et al., 2010; John & Soto, 2007; Braun, Chudowsky, Koenig, et al., 2010). No single method for evaluating teachers is inherently preferable or superior on its own; each has "advantages and limitations" (Peterson, 2000, p. 91), and each "contributes evidence to making a larger case for teacher quality" (Peterson, 2000, p. 91). In the case of teacher evaluation, multiple measures are expected to provide a more complete picture of teacher performance (Goe, Holdheide, & Miller, 2011); finer, more stable categories for classifying teachers (DePascale, 2012; Steele, Hamilton, & Stecher, 2010); feedback to help improve classroom practice (Duncan, 2012); reduced incentives for gaming the system (Steele et al., 2010); and greater confidence in results among stakeholders

(Glazerman et al., 2010).

A variety of measures have been proposed for collecting information about teacher effectiveness, each with a distinct set strengths and limitations for measuring different aspects of this complex construct. Broadly speaking, these measures are often classified as either measures of professional or instructional practice, or measures of student growth and achievement (Partee, 2012). Peterson (2000) lists student reports (e.g., the Tripod Assessment of Instructional Quality and Student Engagement (Ferguson, 2010)), peer reviews, student achievement data, teacher tests (e.g., the Mathematical Knowledge for Teaching Questionnaire (Hill, Schilling, & Ball, 2004), parent reports, documentation of professional activity, systematic observation (e.g., the Classroom Assessment Scoring System (Pianta, La Paro, & Hamre, 2008)), and administrator reports as types of measures that may comprise a teacher evaluation system. Instructional artifacts and portfolios have been proposed, as well (Crosson et al., 2006; Delandshere & Petrosky, 2010; Martínez, Borko, & Stecher, 2012).

Across state and local education agency plans for teacher evaluations systems, the two most widely used measures are value added model (VAM) scores and classroom observations, although student ratings are gaining traction as a viable third option (Bill and Melinda Gates Foundation, 2010). VAMs are used (or proposed for use) in some form in teacher evaluation systems around the U.S. including some of the largest districts in the country in New York, Los Angeles, Chicago, and Denver, and at the state level in Tennessee, District of Columbia, Louisiana, Missouri, North Carolina, Ohio, and South Carolina, among others (Partee, 2012). Classroom observation is a central part of nearly every teacher evaluation system—enacted or proposed—nationwide (e.g., Hill, Charalambous, & Kraft, 2012; Hill & Grossman, 2013). Martínez, Taut, and Schaaf (2013) noted that "indeed, all states recently granted funding under the new Race to the Top legislation in the United States included a new or redesigned classroom observation

component for teacher evaluation" (p. 6). Because these measures are so widely used, VAMs and classroom observations are discussed in more detail in below, prior to discussing the use of student surveys.

Before discussing these measures, however, it is worth providing more detail about the statement that "sound teacher evaluation systems must rely on multiple complementary indicators in order to be valid, comprehensive, and useful." Similar language appears in Race to the Top legislation, where teacher evaluation systems are defined as using "multiple valid measures in determining performance levels" (U.S. Department of Education, 2012, p. 21). Popham (1997, 2013) offers the reminder that validity is a property of inferences, and not a property of measures (APA, NCME, AERA, 1999). Relatedly, it is worth noting that the scores produced in evaluation systems (proposed or enacted) by state and local education agencies have many intended uses (Brandt, 1995; Peterson, 2000; Popham, 2013). Some of these uses are formative. Others, summative. Teacher evaluation systems are used, for example, to identify struggling teachers for assistance, remediation, sanction; or dismissal; to offer incentives to higher performing teachers, including through merit pay or pay-for-performance programs; to inform school practice and district policy on teacher professional development; and developing models of effective instruction to scale up to classrooms across the system (e.g., Millman & Darling-Hammond, 1990). When considering whether the inferences about teacher quality based on measures included in teacher evaluations system are valid, it is also important to consider how these scores are to be used (Messick, 1989; Shepard, 1997), and that different score uses may necessitate different validity arguments.

1.2.1 Value added models and teacher evaluation

Value added models (VAMs) purport to isolate and estimate teacher contributions to student achievement, and have received considerable attention both in the research community and in the media (e.g., Ewing, 2011; Rothstein, 2009; Song &

Felch, 2011). Much has been written about the potential conceptual, statistical, and practical issues complicating the use of VAMs in teacher effectiveness research and policy. There are statistical issues around causal attribution and the proper specification of a counterfactual (Rubin, Stuart, & Zanutto, 2004), the influence of the non-random sorting of students into classrooms and teachers into schools (Braun, 2004, 2005; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Rothstein, 2009), how to appropriately model the persistence of teacher effects (Braun, 2005), how to deal with missing data (Amrein-Beardsley, 2008; Lockwood, Doran, & McCaffrey, 2003), the tenability of the linear mixed-model (Braun, 2004), and the construct validity of the test scores upon which the VAM scores are based (Reckase, 2004). There are also issues with the stability (Papay, 2011) and reliability (Harris, 2009; T. J. Kane & Staiger, 2001) of value added estimates and their sensitivity to the particular assessments they are based upon (Lockwood et al., 2007).

There are other issues with the use of VAMs in teacher evaluation, particularly if they are to be used for formative assessment. VAM scores lack diagnostic value (e.g., Goe et al., 2011) and "cannot produce direct evidence about the effectiveness of educational practices." (Raudenbush, 2004, p. 12). Rothstein and Mathis (2013) noted that there is little evidence that value added scores can be used to provide feedback to teachers to improve instruction, as this would require "texture about the areas in which a teacher is performing well or badly. It is not at all clear that value-added scores—which amount to a single number—can be used for this kind of formative purpose." (p. 11) A robust and credible teacher evaluation system would "examine what teachers actually do in the light of best practices", and "provide constructive feedback to enable improvement" (Haertel, 2013, p. 26). In other words, a key component of a good teacher evaluation system—one that can be used for both teacher improvement and personnel decisions (Haertel, 2013, p. 25)—is a measure of a teacher's instructional practice.

1.2.2 Classroom observations and teacher evaluation

Classroom observations are often seen as an indispensible method to measure instructional practice. In fact, classroom observation is widely regarded as the gold standard for data collection in research on teaching. (Rowan & Correnti, 2009). Many states and districts (for example, Alaska, Arizona, Delaware, District of Columbia, Mississippi, Maryland, New York, South Carolina, among others (Partee, 2012)) rely heavily on observation to provide information about teacher practice, and to identify areas in need of improvement to inform feedback and professional development (Pianta & Hamre, 2009). This has been bolstered by recent work (e.g., Taylor & Tyler, 2012) that has demonstrated that teacher evaluation based on rigorous classroom observation can improve teacher practice. However, observations also have many limitations. First and foremost, they are expensive (Rothstein & Mathis, 2013; Rowan & Correnti, 2009). As pointed out by Balch (2012), in large districts, it is possible that a thoughtful and thorough observation system can translate into full-time positions for dozens of employees, at the cost of several millions of dollars per year. Adding to this, recent studies (Hill et al., 2012; Ho & Kane, 2013) have suggested that obtaining reliable scores from student surveys may pose significant administrative challenges. Additionally, observations may not capture "potentially important dimensions of effectiveness" (Rothstein & Mathis, 2013, p. 9), particularly those that are content-specific (Hill & Grossman, 2013). Many observation protocols are designed to be used across a broad range of classrooms, subjects, and grade levels. However, there is a long research tradition (Lampert, 2001; Shulman, 1987) that shows that the quality of a teacher's instruction is highly content specific. Additionally, recent work (T. J. Kane, McCaffrey, Miller, & Staiger, 2013; Mihaly, McCaffrey, Staiger, & Lockwood, 2013) has shown that correlations between classroom observation ratings and VAM scores are quite low. As noted by Hill and Grossman (2013):

> If observation and student-test-based scores diverge in new systems within states and districts, then teachers might receive conflicting messages about improvement; the feedback on their own instruction may not relate to what they are incented to do based on student test scores. For example, teachers might receive high scores from value-added models but low scores on observation measures . . . This constitutes a major problem that policy makers may grapple with over the next few years. (p. 377-378)

1.2.3 Student surveys and teacher evaluation

Partly in response to the limitations of classroom observations, student surveys of instructional practice have gained support among researchers and policymakers as a viable, cost-effective alternative to traditional observation. In fact, student ratings are one of the oldest available methods for measuring instructional practice. Both Good and Mulryan (1990) and Follman (1992) noted that the Kratz (1896) study of teacher quality in Sioux City, Iowa, was based on student ratings. Peterson (2000) recommended the use of student ratings for the purposes of teacher evaluation. However, a range of methodological and conceptual complications also arise when using student surveys of instructional practice. Historically, concerns have been raised regarding student bias (e.g., Bush, 1954; Peterson, Wahlquist, & Bone, 2000) halo-effects, and whether student ratings are "opinion polls, not teacher evaluations" (Oldham, 1974, in Eastridge, 1976, p. 52)

But there is another set of "thorny methodological issues" (Popham, 2013, p. 9) that arise because of how teacher ratings are derived from student ratings. Specifically, when they are used to generate indicators of a teacher's instructional practices, student surveys explicitly or implicitly assume a specific multilevel measurement model, with individual students nested within classroom or teachers. Chan (1998) noted that the validity of inferences about organizational properties

(such as classroom climate, instructional practice or teacher quality) resulting from aggregated variables is complex and has "not been addressed adequately" (p. 234) in the research literature. In the specific context of student ratings, one key question is what dimensions of instructional practice that are discernible based on aggregated student ratings (e.g., Popham, 2013). The primary analytic tool used to assess the dimensionality of survey instruments is factor analysis (e.g., John & Soto, 2007). However, conventional factor analysis may yield biased or distorted inferences when applied to data that is hierarchical in structure (e.g., Julian, 2001; Schweig, 2013; Zyphur, Kaplan, & Christian, 2008). While this issue has received some attention in methodological literature, there are many unresolved issues in the application of factor analytic techniques to multilevel data, particularly in the context of student surveys. There is, in fact, a consensus that the issues raised by applying factor analytic techniques to multilevel data is an area in need of more research and investigation (Lüdtke, Robitzsch, Trautwein, & Kunter, 2009; Marsh et al., 2012; Sirotnik, 1980).

1.3 Contribution of the current study

This study proposes a new approach for exploring the dimensionality of aggregated student ratings, and collecting validity evidence to support the use of student surveys as indicators of teacher quality. This approach uses a cluster bootstrap and asymptotic distribution free (ADF) theory (Browne, 1974, 1982, 1984) to develop a framework for level-specific model evaluation as described in Yuan and Bentler (2007). This method allows for the explicit testing of measurement hypotheses and psychometric evaluation of aggregates even when the data contains multiple layers of nesting (for example, if students are nested in classrooms, and classrooms are nested in teachers, and teachers are nested in schools, and the unit of analysis is the classroom). The utility of this new approach is illustrated by examining the

dimensions of instructional practice that are discernible based on a student survey that was piloted for potential use as an indicator of teacher quality in the state of New Mexico.

The remainder of this dissertation is structured as follows. Chapter 2 provides a review of the policy literature on using student surveys as a measure of teacher quality or instructional practice, and (re)introduces, in more specific terms, the conceptual and methodological challenges raised by these surveys. Chapter 3 presents conceptual foundations of multilevel factor analysis along with the most commonly used statistical models and test statistics. Chapter 4 introduces the cluster bootstrap, method and describes its application to multilevel factor analysis. Study methods are detailed in Chapter 5. Results from a series of simulation studies investigating the performance of various test statistics, parameter bias and variability, and the performance of standard errors in both two-level and three-level hierarchically structured data sets are presented in Chapters 6. Chapter 7 demonstrates the use of the cluster bootstrap techniques on a real-world data set, based on a statewide survey of instructional practice. Chapter 8 offers a discussion of the results and their implication for researchers and policymakers, and suggestions for future research.

CHAPTER 2

Student Surveys of Instructional Practice

This chapter reviews the existing literature on using student surveys in teacher evaluation. In particular, this chapter discusses: 1) Literature that describes the arguments in favor of including student surveys in teacher evaluation; 2) Current ways in which states and local education agencies have incorporated student surveys into teacher evaluation systems; 3) Some of the conceptual and methodological challenges presented by student surveys; 4) Ways in which current research has addressed (or failed to address) those conceptual and methodological challenges.

2.1 The argument for using student surveys in teacher evaluation

Proponents of student surveys cite several key reasons for including student surveys as measures of instructional practice. These reasons were summarized in Burniske and Meibaum (2011) and include 1) Students themselves are natural observers of the classrooms in which they work and study; 2) Student ratings have proven to reliably discriminate between teachers; 3) Aggregated student ratings are not influenced by rater demographics; 4) Student ratings of instructional practice show relatively robust correlations with student achievement. The following section provides some additional detail about each of these arguments.

2.1.1 Students as natural observers

As Ferguson (2012) stated, students "spend hundreds more hours in each classroom than any observer ever will." (p. 24) and they "know good instruction when they experience it as well as when they do not" (p. 28). Similarly, (Follman, 1992) noted that students were well qualified to serve as raters of their teachers because "no other individual or group has their breadth, depth, or length of experience with the teacher" (p. 169). Worrell and Kuterbach (2001) noted that "it is perhaps not surprising that students can provide accurate ratings of teacher behavior as students spend as much time observing their teachers as their teachers spend observing them" (p. 245). Veldman and Peck (1969) noted that student observations "are the product of observing the teacher on many occasions under normal conditions, and hence avoid many of the obvious problems encountered in typical "one-shot" classroom observations" (p. 107). From an administrative perspective, this line of reasoning makes student surveys an attractive alternative (or complement) to traditional observation. Since the most resource-intensive components of a rigorous observation system involve rater training (e.g., Hill & Grossman, 2013) and deploying multiple raters to classrooms on multiple occasions (e.g., Ho & Kane, 2013) in order to obtain reliable teacher scores, positioning students as raters potentially alleviates this tremendous administrative burden.

2.1.2 Student ratings are reliable

Recent results (e.g., Bill and Melinda Gates Foundation, 2010; Ferguson, 2010) have shown that aggregated student ratings and can be used to reliably distinguish between the practices of different teachers. For example, Ferguson (2010) demonstrated that, there are "massive" (p. 7) disparities in how strongly students endorse items about classroom climate and instructional practice. The average ratings for classrooms in the top decile were nearly four times higher than the

average ratings for classrooms below the bottom decile on some scales. Ferguson (2010) noted, "imagine how different life must be in these ... segments of the classroom quality distribution!" (p. 7).

Other studies have reached similar conclusions. Balch (2012) determined that student surveys achieved adequate internal consistency, and reported scale reliabilities between .704 and .893 (p. 45) for a student survey administered to elementary school students in the state of Georgia. Peterson et al. (2000) reported reliability coefficients of between .76 and .92 (p. 146) for primary, elementary, and secondary student surveys administered in Utah. Peterson et al. (2000) concluded that student surveys can be a reliable and valid data source for teacher evaluation. Follman (1992) summarized the history of reliability studies on student surveys, and found that "a 70-year overview of these ... studies, more than 20 of them, indicates clearly that secondary level student raters— just as older, adult raters— have and can rate teachers reliably, including chance halves, concordance, split half, interclass, internal consistency, and most importantly, stability measures" (p. 170-171).

2.1.3 Sensitivity to student demographics

One potential issue with student surveys is similar in nature to one issue raised about VAMs. That is, students are systematically sorted into particular schools and particular classrooms, and it may be that the student ratings reflect background characteristics of the students, and are not objective ratings of a teacher practice or classroom climate. Veldman and Peck (1969) administered the Pupil Observation Survey to secondary school students in Texas, and found that the scores "are not badly biased by such aspects of the context as the grade level of the class or the socio-economic level of the school" (p. 107). The authors noted, however, that subject matter has a "powerful influence" on the scores (p. 107). Thompson (1974) noted:

> Data from several high school studies have been analyzed by the author to determine if responses have been significantly affected by student characteristics, including: sex, year in school, grade point average, expected course grade, hours spent studying, and absenteeism. No significant relationship has been found between these items and student rating of faculty performance. (p. 26)

2.1.4 Robust correlations with student achievement

There is an increasing body of research demonstrating that the ratings of instructional practice produced by student surveys are predictive of student achievement. Recent work has shown that indicators derived from student surveys correlate significantly with VAM scores (Bill and Melinda Gates Foundation, 2010; Mihaly et al., 2013). A study by Wilkerson, Manatt, Rogers, and Maughan (2000) based in a school district of Wyoming found that student ratings of teachers are more strongly correlated with achievement than either principal ratings or teacher self-ratings. Those results held across elementary, middle, and high school students. Those authors concluded that "student ratings of teachers are the best predictors of student achievement among groups of raters when the focus is student performance" (p. 190). Similar results were found by Kyriakides (2005), based on data collected in Cypress.

2.1.5 A tradition of use in higher education

There is a long tradition of using student surveys in higher education. Marsh (1987) provided an extensive overview of this history, and notes that student evaluation programs were introduced at Harvard, the University of Washington, Purdue University and the University of Texas and other institutions in the mid-1920s (p. 257). Marsh (1987) noted that "the term 'students' evaluations of teacher

performance' was first introduced in the ERIC system in 1976; between 1976 and 1984 there were 1055 published and unpublished studies under this heading" (p. 257). Many studies have found that, at the university level, student evaluations can provide reliable scores and valid inferences about instructional practice (e.g., Aleamoni, 1999; Feldman, 1978; Marsh, 1987; Toland & De Ayala, 2005). While many aspects of secondary and elementary schooling are clearly different from those encountered in university settings, many authors that have written about the use of student surveys in K-12 educational settings have drawn heavily from the research base in higher education. For example, the research syntheses in Aleamoni (1987, 1999) is referenced by both Follman (1992) and Peterson et al. (2000). Follman (1992) noted:

> Since, logically there is some justification in viewing high school students as a downward extension of college students, perhaps quantitatively but not qualitatively different, some authorities view rating instructors at the high school level as similar to rating instructors at the college level (p. 174).

Burniske and Meibaum (2011) also referenced Aleamoni (1999) to provide a historical context for the use of use of student surveys in teacher evaluation systems. Burniske and Meibaum (2011) did not note that Aleamoni (1999) was concerned with university-level student ratings, rather than those that would be used in K-12 educational settings.

2.2 State and local school districts and student surveys

Student surveys are increasingly prominent in district plans for teacher evaluation. Table 2.1 provides information about the states and several notable local school districts that have included student surveys in their teacher evaluation plans. Not all of the states have articulated specific weights or survey instruments, and so

some cells in this table are left blank. There are many plans that specifically call for

Table 2.1: Student surveys and teacher evaluation

States or local districts specifically including student surveys		
Name	Survey instrument	Weight in overall evaluation
Hawaii	Tripod Survey	10%
Georgia	My Student Survey	10%
Maine	Tripod Survey	10%
Massachusetts		
Minnesota		15%
Kentucky	Student Voice Survey	
Chicago, IL		10%
Denver, CO	Tripod Survey	5%
Memphis, TN	Tripod Survey	5%
New York City, NY	Tripod Survey	5%
States or local districts where surveys are optional measures		
Name	Survey instrument	Weight in overall evaluation
Arizona		17%
Alaska		
Colorado		Contributes to 50%
Connecticut		5%
Idaho		Contributes to 67%
Michigan		Contributes to 20%
New Mexico		Contributes to 25%
New York		Contributes to 60%
States or local districts exploring the use of student surveys		
Name	Survey instrument	Weight in overall evaluation
North Carolina	Tripod Survey	
North Dakota		
South Dakota		
Washington		

student surveys to be included in teacher evaluations. These include six states and four local districts. In most of these districts, student surveys count for between 5 and 10% of a teachers overall evaluation. Seven other states mention student surveys as a possible measure of teacher quality. In these states, student surveys are typically described as one possible option for inclusion in a teacher evaluation system, along with other potential measures including parent surveys and portfolios of student work or other classroom artifacts, and the relative contribution that the student survey makes to the overall teacher evaluation varies considerably. In Connecticut, student surveys or indicators of school-level learning (such as a School Performance Index) or a combination of the two may be used in teacher evaluation,

and represent 5% of the total summative evaluation score (Connecticut State Department of Education, 2012). In Michigan, the state recommended that local education agencies "may use other data that provide evidence about a teacher's practice (e.g., student surveys, parent surveys, portfolios), but for no more than 20% of the practice section" (Michigan Council for Educator Effectiveness, 2013, p. 11). In Colorado, student surveys are listed as one possible measure of instructional practice that can be used in conjunction with classroom observations. Measures of professional or instructional practice account for half of a teacher's evaluation (Colorado Department of Education, 2013).

Lastly, several states and local districts are in the process of exploring the potential for student surveys to be used in teacher evaluation. North Carolina piloted the Tripod Project survey as a part of that state's teacher evaluation system in the spring of 2012 (Partee, 2012). Similar surveys are in use or being piloted in Los Angeles (Phillips & Yamashiro, 2013).

There are two interesting trends in how states and local districts have incorporated student surveys into their teacher evaluation plans. Overwhelmingly, states and local districts have turned to the Tripod Survey (Ferguson, 2010) as the survey instrument of choice. Of the 8 states and local districts that mention a specific survey instrument, 6 of them have elected to use the Tripod Survey specifically. Secondly, while many education agencies have included student surveys as a measure of instructional practice (e.g., Colorado and New York), others have positioned student surveys as a secondary measure of student outcomes (e.g., Connecticut). In this way, positive engagement, for example, is seen as a student outcome, rather than as a measure of a teacher's practice.

2.3 Conceptual issues with surveys of teacher practice: the unit of analysis

The combination of utility and administrative feasibility make student surveys attractive as a measure of instructional practice. However, there remain some critical questions about the psychometric properties of the scales measured by student surveys of instructional practice. Lüdtke et al. (2009) stated that "there are serious conceptual and methodological challenges that need to be addressed before student ratings can properly be used to gauge the effects of characteristics of the learning environment" (p. 120-121). Specifically, one of the most pressing conceptual issues concerns the unit of analysis (Lüdtke et al., 2009) of the survey. In other words, researchers should pay careful attention to whether the survey designed to measure individual perceptions, or classroom level phenomena. Guion (1973) described this fundamental conceptual distinction as a question of whether the survey is concerned with individual attributes, or with attributes of the world that individual inhabits.

This conceptual complication is inherently connected to the way in which students are organized into classrooms. When surveys are administered to students, individual students are grouped (or clustered) within specific classrooms. This creates what is commonly referred to as a hierarchical (Goldstein, 2003; Hox, 2010; Raudenbush & Bryk, 2002) system. It may also be said that students are nested within teachers (e.g., Brennan, 2001; Shavelson & Webb, 1991). In this tradition, throughout this paper, students are often referred to as either the student level, the within-group level, or as level-1 units, and teachers are referred to as either the teacher level, the between-group level, or as level-2 units. Because students are nested within classrooms, survey responses may be used in two distinct ways, and there can be (potentially) two levels of analysis that are of substantive interest. First, the student responses can be used to analyze the individual perceptions of

students. Second, the individual student responses can be aggregated "to yield a measure of the "shared perception of the environment"" (Lüdtke et al., 2009, p. 121). In other words, the means of the level-1 variables can be used as a measure of a level-2 phenomenon. This presents researchers with a set of decisions that do not exist in conventional research contexts, where the data is not grouped or clustered. Lüdtke et al. (2009) described an example of how researchers may be confronted with this issue:

> If a researcher is interested in the effects of a supportive class climate on student motivation, is it appropriate to examine the relationship between a students individual perception of classroom climate and his or her motivation? Or would it make more sense to aggregate student perceptions at the classroom level and to analyze the association of the aggregated score with the outcome variables? (p. 121)

Sirotnik (1980) examined the issue of unit of analysis by presenting a set of hypothetical items that could be asked of teachers about whether their schools offered trusting environments. Again, there is a hierarchical system here, with teachers nested within schools. There are two levels that can be considered here, as was the case with the student survey example from Lüdtke et al. (2009). Individual teacher responses can be used to analyze the perceptions of teachers. Or, teacher responses can be aggregated to yield a measure of the school environment. Sirotnik (1980) asked:

> are [we] measuring something about the teacher, something about the school, or both? Suppose item means are computed. Do these means represent measures of an intrinsic property of the school, with averaging over teachers being incidental and merely a convenient operational device for getting at this property? Or do these means represent the central tendency of distributions of measures of a property of teachers?

(p. 261)

Sirotnik (1980) and Lüdtke et al. (2009) were not the only researchers to raise this concern. The unit of analysis concern has a lengthy tradition in the social sciences, and the seminal work of Robinson (1950) and Alker (1969), as well as similarly influenced work in the organizational management literature. Glick (1985); James (1982); Roberts, Hulin, and Rousseau (1978) raised similar concerns about the issues that can arise when researchers do not pay careful attention to issues involving the unit of analysis.

The question raised by Sirotnik (1980), regarding the possibility that survey items are measuring something about level-1, level-2, or both levels at the same time presents a particular conceptual complexity. It may be necessary for researchers to develop two different sets of theories about the relationship of survey items to underlying theoretical constructs, one for the individual level, and one for the group level. Glick (1985) cautioned researchers not to assume that "organizational and psychological climate have the same dimensionality and the same pattern of relationships with variables of interest" (p. 605). In other words, it is problematic to merely assume that the same number of variables, and same patterns of relationships exist between aggregated survey variables (those that refer to between-group phenomena) and individual level variables (those that refer to individual perceptions). Glick (1985) continued to say that between classroom relationships should be "empirically confirmed or disconfirmed, not definitionally asserted" (p. 605). Guion (1973) noted that individual perceptions of climate and aggregated climate variables may have "distinguishably different networks of correlates." (p. 122).

If the unit of analysis is the classroom, a decision about whether student level phenomenon are of substantive interest should be based, in part, on considerations of whether differences between student responses represent meaningful differences in the true standing of students, or whether differences between student responses

represent measurement error. Marsh et al. (2012) outlined this distinction in detail and refer to this as the distinction between "context" and "climate" variables. Context variables have, as their reference, the individual student, and differences between students are substantively meaningful. Classroom averages can be used to describe classroom composition, but these averages essentially represent "the central tendency of a distribution of measures" in the sense described by Sirotnik (1980). Individual students are not exchangeable. Marsh et al. (2012) used gender as an example of a context variable. Each student has a gender, and that gender is not exchangeable with other students in the class. A class can have a gender composition (proportion male), and that gender composition may be importantly related to other individual or classroom level phenomena.

On the other hand, climate variables have as their reference, the teacher or classroom level. In this model, it is assumed that every student has the same (or at least very similar) mental image of classroom climate or instructional practice, and that this common conception contributes to common variance among item responses. Thus, with climate variables, the student responses are fundamentally exchangeable, and the variance between students within the same classroom is attributed to sampling error and represents "noise", while variance in student aggregate ratings between classrooms is assumed to represent true variance in classroom quality. The aggregated ratings measures of an intrinsic property of the group level, in the sense of Sirotnik (1980). In the organizational management literature, such aggregated variables are commonly referred to as composition or reflective aggregation variables (Bliese, 2000; Chan, 1998; Kozlowski & Klein, 2000; Lüdtke et al., 2009).

As measures of instructional practice, student surveys are fundamentally concerned with the aggregated (also called the between-classroom or between-teacher) level. Students are positioned as raters of the classrooms in which they study, and in this way, student ratings are exchangeable, and the variation between students

represents noise. The variables derived from student surveys are, thus, most commonly conceived of as climate variables in the sense defined by Marsh et al. (2012) As Lüdtke et al. (2009) described:

> The main purpose of collecting individual students' ratings of their class ... is to assess aspects of environments that are clearly located at the group level (e.g., class level). Thus, students are regarded as informants on their learning environment, in the sense of multiple observers providing data on one construct. At the individual level, the measurements refer to the phenomenology of the student. At the class level, however, they refer to differences between classes. If educational researchers want to gain insights into the effects of learning environments, they have no choice but to use aggregated student data (p. 121).

Because the primary unit of analysis for student survey is the classroom or the teacher, understanding the constructs measured by the aggregated survey responses is critical for developing and testing theories about how aspects of a teachers instructional practice influence other variables of substantive interest, such as student achievement and persistence in school.

2.4 Understanding the constructs measured by the aggregated survey responses

One way to assess the psychometric properties of aggregated survey variables, and to understand the constructs measured by aggregated survey responses is to use multilevel factor analysis. In a multilevel factor analysis, covariance among the items is caused by unobserved (latent) differences among individual students, and covariance between aggregated item scores is caused by unobserved (latent) differ-

ences among classrooms/teachers. Reiterating the recommendations of Cronbach (1976) and Longford and Muthén (1992), different sets of factors can be tested on the level-1 units and the level-2 units. The theoretical and statistical framework for multilevel factor analysis is discussed in more detail in Chapter 3. Marsh et al. (2012) encouraged researchers to use multilevel factor analysis to test the between-level factor structure explicitly (p. 122). In fact, when the variables are conceived of as climate variables, it is often the case that the between-level factor structure is the only structure that is relevant (Marsh et al., 2012, p. 122), and that the within-level factor structure is not substantively interpretable. Echoing the type of concern raised by Glick (1985) and Guion (1973), Marsh et al. (2012) also cautioned against assuming that the factor structure is necessarily the same at the individual student level and at the classroom level. In fact, factor analytic research dating back to Cronbach (1976) cautioned that a researcher might need, "one set of factors for his between-groups theory and another set of factors for his within-groups theory. To be sure, he may find that the two sets of constructs coincide, but that is a possibility to be evaluated, not assumed" (p. 203). Longford and Muthén (1992) also noted that a between-groups phenomenon may be completely unrelated to a within-groups phenomenon.

Other notable sources have also recommended the use of multilevel factor analysis procedures to investigate the measurement structure of level-1 phenomena, level-2 phenomena, or both. Several of these investigations found evidence that different sets of factors could be found at different levels of analysis. Härnqvist (1978) decomposed student ability test scores into individual, class, and district components, and studied the factorial structure of the within and between class variables separately. Härnqvist (1978) found 5 within-class ability factors, and 2 between-class factors, providing an empirical illustration that different sets of factors may be necessary at different levels of analysis. Holfve-Sabel and Gustafsson (2005) analyzed results from a student survey and found evidence for seven

within-classroom factors and three between-classroom factors. Other recent research (Reise, Ventura, Nuechterlein, & Kim, 2005; Zyphur et al., 2008) has shown that patterns of factor loadings may also differ across different levels of analysis, even if the number of factors is the same.

There are also several studies that have found the same factor configurations at both levels of analysis. D'haenens, Van Damme, and Onghena (2010, 2012) use multilevel exploratory and confirmatory factor analysis to investigate school climate variables, and found the same number of factors at each level, with the same patterns of factor loadings. In an investigation of a university level student survey, Toland and De Ayala (2005) also found the same number of factors at level-1 and level-2. Fauth, Decristan, Rieser, Klieme, and Büttner (2014) investigated the factorial structure of a primary school survey of teaching quality, and found evidence for the same factorial configuration in student ratings of instructional behavior in mathematics classrooms at the student level and the teacher level. Kunter et al. (2008) also found the same factorial configuration at the student and teacher levels.

2.5 Ignoring the unit of analysis issue: historical approaches to student surveys

Despite the fact that the studies that have explicitly examined level-1 and level-2 constructs have found evidence for a wide range of patterns (i.e., there are cases where number of constructs differ, cases where patterns of factor loadings differ, and cases where there is evidence of configural (Meredith, 1993) invariance), the unit of analysis issues described above are, to a large extent, absent from the literature on classroom climate and a discussion of the conceptual and methodological challenges involved with using student surveys are almost completely absent from state documentation on the student surveys used in teacher evaluation. Lüdtke

et al. (2009) noted explicitly that these issues "have not yet received sufficient research attention" (p. 120-121). Sirotnik (1980) noted that concerns about the unit of analysis during instrument development are "virtually nonexistent" (p. 256). Marsh et al. (2012) noted that, "despite the clear resolution of this methodological issue for more than a quarter of a century," there is still "ongoing confusion in the educational literature" (p. 111) about the appropriate ways to consider the unit of analysis in the study of classroom climate.

Many of the states that have required student surveys to be incorporated into teacher evaluation have provided little written documentation about the psychometric properties of the scores produced by the surveys, and concerns about the unit of analysis are not addressed at all. For example, the state of Hawaii stated only that the Tripod survey is "well designed" (Hawaii State Department of Education, 2013, p. 16). The state documents justify the use of Tripod based on results from the Measuring Effective Teaching (MET) project, a three year study on teaching and teacher quality funded by the Gates Foundation. The Kentucky Department of Education also cited results from the MET project to support their use of student surveys for the purposes of teacher evaluation (Kentucky Department of Education, 2012). New York City, Memphis, and Denver were all school districts that participated in the MET project. However, as Camburn (2012) noted, reports associated with the MET project contain little empirical evidence related to the psychometric properties of the student survey included in that study.

Thus, while the investigations listed above (Cronbach, 1976; D'haenens et al., 2010; D'haenens, Van Damme, & Onghena, 2012; Fauth et al., 2014; Härnqvist, 1978; Holfve-Sabel & Gustafsson, 2005; Kunter et al., 2008; Toland & De Ayala, 2005) may leave the impression that researchers using student surveys of classroom climate or instructional practice have thoughtfully considered the unit of analysis, these studies are the exception, rather than the rule. Much of the literature on student surveys does not fully and properly consider the conceptual issues sur-

rounding the unit of analysis, and often times, the research base contains evidence for the reliability and validity of scores produced by student surveys as if the unit of analysis were individual students, rather than teachers or classrooms.

2.5.1 Reliability and the unit of analysis

Follman (1992) provided a detailed review of the literature on the reliability of aggregated scores produced from student surveys of teacher quality and instructional practice, dating as far back as the early 20th century. Of the twenty studies cited by Follman (1992), nearly all of them investigate the reliability of scores using coefficient alpha, split half reliability, or test-retest reliability. For all of these reliability coefficients, the unit of analysis is the individual student, rather than the aggregated ratings. There are several lines of research (Bliese, 2000; Brennan, 1995; James, 1982) that have shown that the reliability of group means may not be related to the reliability of individual scores in any systematic way. In particular, high reliability of individual scores does not imply high reliability of group means, and vice versa. The study by Wilkerson et al. (2000) reported coefficient alpha (Cronbach, 1951) as an index of reliability, even though the unit of analysis is not individual students. Balch (2012), in a validity study of My Student Survey (Table 2.1), the survey used in Georgia's teacher evaluation system, also reported coefficient alpha.

It should be noted that there is a strong tradition in organizational management of using reliability coefficients based on intraclass correlations (ICC). For example, Bliese (2000), James, Demaree, and Wolf (1984) and Lüdtke et al. (2009) all described using a version of the Shrout and Fleiss (1979) $ICC(1,k)$ in order to ascertain the reliability of group means. In a similar tradition, and informed by work from Generalizability Theory (Cronbach, Gleser, & Nanda, 1972), M. T. Kane and Brennan (1977) and O'Brien (1990) employed multilevel measurement models to explore the reliability of classroom level measures at the appropriate level of

analysis. Raudenbush, Rowan, and Kang (1991) also considered the issue of the reliability of school climate variables at the appropriate unit of analysis. Reliability studies for the Tripod Survey (Ferguson, 2010) explored the reliability of classroom level measures at the appropriate level of analysis, as did the work of Kunter et al. (2008). However, the application of multilevel models to the study of reliability in classroom-level variables is "less common" (Raudenbush, Martinez, Bloom, Zhu, & Lin, 2010, p. 8).

2.5.2 Factor analysis and the unit of analysis

In factor analytic investigations of student surveys, the unit of analysis is also commonly ignored. Specifically, researchers often use conventional factor analysis techniques that do not allow for the possibility that, as was noted above (Cronbach, 1976; Glick, 1985; Longford & Muthén, 1992), the aggregated variables have different dimensionality and patterns of relationships than the individual level variables. In particular, conventional factor analysis procedures are often applied either on the disaggregated data, or on the group means (weighted or unweighted). Each of these procedures is described below.

2.5.2.1 Disaggregated factor analysis

States and local districts commonly use conventional factor analysis methods to establish the between-classroom or between-teacher factor structure. The validation study of My Student Survey (Balch, 2012) found validity evidence for inferences regarding aggregated survey variables based on the results of a conventional factor analysis performed on the disaggregated data. Similarly, student surveys of school climate like the South Carolina School Climate survey (DiStefano, Monrad, May, McGuiness, & Dickenson, 2007), the Working Conditions Survey (Moir, 2009), all

used a single-level factor analysis on either the student level data (ignoring the nesting of students within classrooms). Ladd (2011), examined the relationship between teacher working conditions and teacher retention with working conditions variables that were derived from an exploratory factor analysis on the disaggregated correlation matrix. Ryan and Patrick (2001) used a similar approach to investigate the relationship between classroom environment and student motivation and engagement.

There are many problems with such an approach. Firstly, this approach imposes the (rather stringent) assumption of cross-level measurement invariance. For example, in a two-level situation where students are nested within classrooms, the student level and classroom level factor structures are constrained to be equal. However, many of the empirical the examples provided above (Härnqvist, 1978; Holfve-Sabel & Gustafsson, 2005; Reise et al., 2005; Zyphur et al., 2008) show that factor structures can vary considerably across levels, and that analyzing multilevel data with conventional factor analytic techniques can lead to "substantively misleading" inferences about relationships among indicators, or between indicators and external variables (Reise et al., 2005, p.130). Additionally, there are statistical issues that are introduced by using conventional factor analysis procedures on data that is hierarchically structured, and research has shown that this can lead to biased parameters and standard errors and inflated test statistics (e.g., Julian, 2001; Preacher, Zyphur, & Zhang, 2010).

2.5.2.2 Factor analysis on group means

Factor analyses performed on the Tripod Survey are based on classroom aggregates (Ferguson, 2010), as are factor analyses on the New York City Department of Education Environmental Surveys (Rockoff & Speroni, 2008). Worrell and Kuterbach (2001) used classroom aggregates as the basis for a factor analysis, as did Hoy and Clover (1986) in their validation of the Organizational Climate Description

Questionnaire. While the use of conventional factor analysis procedures that ignore the grouping of students into classrooms seems to ignore the unit of analysis concerns raised by Lüdtke et al. (2009) completely, it is less intuitive how using the group means fails to address this concern. Hoy and Clover (1986) were explicit that they use group means to address unit of analysis concerns. Hoy and Clover (1986) noted:

> When [a] property is viewed as fundamentally intrinsic to the group, as it is in school climate, then between-school analysis is most appropriate. Unfortunately, total analysis is most frequently used, or more accurately misused, in studies of organizational climatea more appropriate procedure would have been to aggregate the scoresand then factor analyze the item matrix. (p. 98)

However, there are technical issues with this approach. As pointed out by B. O. Muthén (1994), the sample between level covariance matrix, such as the one employed by Hoy and Clover (1986), Ferguson (2010), and Rockoff and Speroni (2008) contains both within and between variance sources. As such, estimates of the between-level factor structure can often be biased by using such a procedure (B. O. Muthén, 1994; Preacher et al., 2010).

The persistence of conventional factor analysis approaches in research on student surveys suggests that methodologists have not done an adequate job of communicating the consequences of using such approaches, and that practitioners continue to see this issue as a methodological distinction without a substantive difference. There is, however, a strong possibility that the application of single-level factor analysis to hierarchically structured data can lead to obscured or spurious information about prediction and correlation among policy relevant constructs. For example, it could result in identifying the wrong number of factors, or associating items with incorrect factors. For example, items in a classroom environment survey may distinguish two within-class latent variables, such as student engagement and

instructional rigor. However, at the classroom level it could be that the level of engagement and rigor in a classroom tend to strongly co-vary (B. O. Muthén & Asparouhov, 2011) so that there is only one broadly defined latent variable at that level. Assuming cross-level invariance would then blind researchers to the substantial overlap among dimensions of instructional practice.

2.6 Summary

Student surveys potentially offer important information about instructional practice. However, little is known about the psychometric properties of these surveys, and relatively little attention has been paid to the conceptual and methodological challenges that arise when working with survey-based indicators. Specifically, there is a relatively small body of literature that investigates the dimensionality of the teacher level variables, and much more work is needed to examine the distinct dimensions of instructional practice that are discernible based upon aggregated student survey responses. In this study, I propose, test, and demonstrate one possible approach to multilevel factor analysis that can be applied to psychometric investigations of the sort described above. In the following chapter, I provide a brief introduction to model estimation and testing in conventional and multilevel factor analysis. Maximum Likelihood, asymptotic distribution free (ADF) (Browne, 1974, 1982, 1984), and residual-based (Browne, 1982, 1984) test statistics are defined. This background helps to situate the tradition of level-specific model evaluation (e.g., Goldstein, 2003; Longford & Muthén, 1992; Yuan & Bentler, 2007), in the larger landscape of multilevel factor analysis. Both two-level and three-level models are discussed.

CHAPTER 3

The Multilevel Factor Analysis Framework

This chapter first presents the statistical framework for the estimation of parameters and standard errors in conventional confirmatory factor analysis, as well as the statistical development of several test statistics commonly used in the confirmatory factor analysis tradition (e.g., Bock, 1960; Jöreskog, 1969; Lawley & Maxwell, 1973). Next, the framework for multilevel factor analysis is presented, as a generalization of conventional single level methods. The estimation of standard errors and parameters in multilevel factor analysis is discussed, and approaches to model testing in multilevel factor analysis are presented. The complications that arise in multilevel factor analysis, including the potential limitations of several commonly used approaches to multilevel factor are discussed in detail.

3.1 Conventional confirmatory factor analysis and model testing

The conventional factor model (e.g., Bollen, 1989; Jöreskog, 1969) can be expressed

$$y = \lambda\eta + \epsilon \tag{3.1}$$

Where y is a p-variate vector of observed scores measuring η. η is an $m \times 1$ vector of latent variable scores on m factors, assumed to be normally distributed with 0 expectation. λ is a $p \times m$ matrix of factor loadings. ϵ represents a $p \times 1$ vector

of residuals (also called uniquenesses, (e.g., Bollen, 1989)), which are assumed to be identically and independently distributed. This factor model yields the following covariance structure model:

$$\Sigma = \Lambda\Phi\Lambda^T + \Psi \tag{3.2}$$

where Λ is the $m \times m$ matrix of factor loadings described above, Φ is an $m \times m$ matrix of factor covariances, and Ψ is a $p \times p$ diagonal matrix containing unique (residual) variances.

Optimal estimates of model parameters, $\hat{\theta}$ are found by minimizing a discrepancy function, $F[S, \Sigma(\theta)]$, which indicates the discrepancy between the sample covariance matrix, S, and the model implied covariance matrix $\Sigma(\theta)$. In general, given a $p \times p$ population covariance matrix Σ and a p-vector of free parameters θ, a testable null hypothesis can be expressed:

$$H_0 : \Sigma(\theta) = \Sigma \tag{3.3}$$

In other words, the population covariance matrix, Σ, is equal to the model implied covariance matrix, $\Sigma(\theta)$ (Bollen, 1989). The null hypothesis is frequently tested using a test statistic based on the discrepancy function, $F[S, \Sigma(\theta)]$.

The following sections provide details about parameter estimation, test statistics, and standard errors that are derived using normal theory (i.e., maximum likelihood estimation and the likelihood ratio test statistic) and Asymptotic Distribution Free (ADF) theory. Six test statistics are defined, including 1) The ADF test statistics T_{ADF} and T_{CADF}, 2) The likelihood ratio test statistic T_{ML}, 3) The rescaled ikelihood ratio test statistic T_{RML} and 4) The residual-based ADF test statistics T_{RADF} and T_{CRADF}.

3.1.1 Asymptotic Distribution Free theory and estimation

Browne (1974, 1982, 1984) suggested a class of estimators that are based on a generalized least squares discrepancy function. Optimal estimates of model parameters, $\hat{\theta}$, are found by minimizing

$$F_{GLS} = (s - \sigma(\theta))^T W (s - \sigma(\theta)) \tag{3.4}$$

Where $s = vech(S)$ is the half-vectorization of S, and $\sigma(\theta) = vech(\Sigma(\theta))$ is the half-vectorization of $\Sigma(\theta)$. For an $p \times p$ covariance matrix, s and $\sigma(\theta)$ contain $p^* = \frac{((p+1)p)}{2}$ elements. W is a $p^* \times p^*$ weight matrix. Following Browne (1984) (see also (Bentler & Dudgeon, 1996; Foldnes, Foss, & Olsson, 2012)) F_{GLS} is correctly specified for W if:

$$W \xrightarrow{p} \Gamma^{-1} \tag{3.5}$$

Where Γ is given by the asymptotic distribution of $\sqrt{n}(s - \sigma(\theta))$:

$$\sqrt{n}(s - \sigma(\theta)) \xrightarrow{d} N(0, \Gamma) \tag{3.6}$$

Since s and $\sigma(\theta)$ are $p^* \times 1$ vectors, the matrix Γ, which describes the asymptotic variances and covariances of $\sqrt{n}(s - \sigma(\theta))$, is a symmetric positive definite $p^* \times p^*$ matrix. In conventional factor analysis, $\hat{\Gamma}$, a consistent estimate of Γ, can be obtained by calculating fourth-order central sample moments (e.g., Bentler, 2006). In this way, it is possible to let W in Equation 3.4 equal $\hat{\Gamma}^{-1}$ without imposing distributional assumptions on the observed variables. This yields the Asymptotic Distribution Free (ADF) discrepancy function (Browne, 1974, 1982, 1984):

$$F_{ADF} = (s - \sigma(\theta))^T \hat{\Gamma}^{-1} (s - \sigma(\theta)) \tag{3.7}$$

When the model is correctly specified the ADF test statistic T_{ADF} is given by:

$$n\hat{F}_{ADF} \xrightarrow{d} \chi^2_d \tag{3.8}$$

Where $\hat{F}_{ADF} = F[S, \Sigma(\hat{\theta})]$, the minimized value of the discrepancy function, and $n = N - 1$ (one less than the total sample N). The degrees of freedom, d, is given by $d = p^* - q$ (e.g., Bollen, 1989).

The parameter estimates obtained by minimizing Equation 3.7 are consistent, but the rate of convergence can be quite slow (e.g., Foldnes et al., 2012). Relatedly, numerous studies (e.g., Curran, West, & Finch, 1996; Hox, Maas, & Brinkhuis, 2010; Hu, Bentler, & Kano, 1992; B. O. Muthén & Kaplan, 1985, 1992; Powell & Schafer, 2001) have shown that T_{ADF} does not perform well at all but the largest sample sizes. Specifically, T_{ADF} tends to over-reject correct models. For this reason, Yuan and Bentler (1997a) suggested a small sample correction to T_{ADF} . The corrected ADF statistic can be expressed:

$$T_{CADF} = \frac{T_{ADF}}{1 + \frac{T_{ADF}}{n}} \tag{3.9}$$

Neither T_{ADF} nor T_{CADF} will be estimable unless $\hat{\Gamma}$ is invertible. Practically speaking, this requires $n > p^*$ (Yuan & Bentler, 1998).

ADF-theory standard errors for estimated model parameters can be obtained based on:

$$Cov\left(\hat{\theta}\right) = \frac{\left[\dot{\sigma}(\hat{\theta})^T \hat{\Gamma}^{-1} \dot{\sigma}(\hat{\theta})\right]^{-1}}{N} \tag{3.10}$$

where $\dot{\sigma}(\hat{\theta})$ is the derivative of $\sigma(\theta)$ with respect to θ, evaluated at $\hat{\theta}$, the estimated model parameters. As is the case with test statistics and parameter estimates based on Equation 3.7, research has shown that standard errors based on Equation 3.10 converge slowly, and cannot be trusted at samples of $N = 1000$ or less. (Curran et

al., 1996, p. 198). Similar conclusions were reached by Hoogland (1999). Yuan and Bentler (1997b) proposed ADF standard errors based on the corrected covariance matrix:

$$\hat{\Gamma}_c^{-1} = \frac{n\hat{\Gamma}^{-1}}{n - p^* - 1} \tag{3.11}$$

3.1.2 Maximum likelihood estimation

The maximum likelihood (ML) discrepancy function (Jöreskog, 1967) is derived from the normal-theory log-likelihood. Optimal estimates of model parameters, $\hat{\theta}$, are found by minimizing

$$F_{ML} = log|\Sigma(\theta)| + tr[S\Sigma(\theta)^{-1}] - log|S| - p \tag{3.12}$$

where $|\cdot|$ denotes the determinant, and tr denotes the trace of a matrix. This discrepancy function can be used to define the following ikelihood ratio test statistic:

$$T_{ML} = n\hat{F}_{ML} \tag{3.13}$$

Results in Browne (1974) showed that the maximum likelihood discrepancy function in Equation 3.12 can be understood as a special member of the class of generalized least squares estimators given in Equation 3.4 with a weight matrix given by:

$$W_{ML} = .5D_p^T[\Sigma(\hat{\theta})^{-1} \otimes \Sigma(\hat{\theta})^{-1}]D_p \tag{3.14}$$

The matrix D_p is a $p^2 \times p^*$ duplication matrix (Magnus & Neudecker, 1988). Under the assumption of multivariate normality, W_{ML} satisfies Equation 3.5. This means that F_{ML} is asymptotically optimal (Browne, 1974; Foldnes et al., 2012), and the parameter estimates $\hat{\theta}$ will asymptotically have minimum variance. Additionally, T_{ML} will be asymptotically distributed as a central chi-square variate when the model is correctly specified. In fact, Browne (1984) suggested that under some

conditions, the weight matrix given in Equation 3.14 may still be correctly specified provided there is no excess multivariate kurtosis in the observed variables.
ML-based standard errors for are given by

$$Cov\left(\hat{\theta}\right) = \frac{\left[\dot{\sigma}(\hat{\theta})^T W_{ML} \dot{\sigma}(\hat{\theta})\right]^{-1}}{N} \tag{3.15}$$

(e.g., Bentler, 2006; Yuan & Hayashi, 2006). Equation 3.15 forms the basis for the computation of standard errors in most covariance structure software. These standard errors are consistent provided the model is correctly specified and there is no excess kurtosis in the observed variables (e.g. Browne, 1984; Yuan & Hayashi, 2006).

3.1.3 The Satorra-Bentler rescaled test statistic T_{RML} and robust standard errors

The Satorra and Bentler (1988) rescaled test statistic T_{RML} is a corrected version of the likelihood ratio statistic T_{ML}. T_{RML} was designed to rescale T_{ML} based on excess skew and kurtosis in the observed variables (Satorra & Bentler, 1988), since T_{ML} is derived under the assumption of multivariate normality, and may not be asymptotically distributed as a central chi-square variate when there is excess kurtosis (e.g., Browne, 1984). Let

$$\hat{U} = W_{ML} - W_{ML}\dot{\sigma}(\hat{\theta}) \left(\dot{\sigma}(\hat{\theta})^T W_{ML} \dot{\sigma}(\hat{\theta})\right)^{-1} \dot{\sigma}(\hat{\theta})^T W_{ML} \tag{3.16}$$

Also let $k = \frac{tr(\hat{U}\hat{\Gamma})}{d}$, where $d = p^* - q$ as above. Then:

$$T_{RML} = \frac{T_{ML}}{k} \tag{3.17}$$

While T_{RML} is not generally chi-square distributed, its expectation is asymptotically equal to the expectation of χ^2_d (e.g. Bentler & Yuan, 1999).

Under normal theory, when the model is correctly specified, the standard errors given by Equation 3.15 are consistent. However, when the distributional assumptions are violated, and particularly when there is excess kurtosis in the observed variables, these standard errors will tend to be biased and consistency can no longer be assumed (e.g. Yuan & Hayashi, 2006). In this case, robust standard errors (e.g., Bentler & Dijkstra, 1985; Shapiro, 1986) can be estimated, based on a triple-sandwich estimator (Huber, 1967; White, 1980):

$$Cov\left(\hat{\theta}\right) = \frac{\left[\dot{\sigma}(\hat{\theta})^T W_{ML} \dot{\sigma}(\hat{\theta})\right]^{-1} \left[\dot{\sigma}(\hat{\theta})^T W_{ML} \hat{\Gamma} W_{ML} \dot{\sigma}(\hat{\theta})\right] \left[\dot{\sigma}(\hat{\theta})^T W_{ML} \dot{\sigma}(\hat{\theta})\right]^{-1}}{N} \tag{3.18}$$

These standard errors are consistent when the model is correctly specified (Yuan & Hayashi, 2006).

3.1.4 The residual-based ADF test statistics T_{RADF} and T_{CRADF}

Browne (1982, 1984) described a second class of test statistics, which can be obtained in conjunction with a wide range of estimators, including ML. The residual based test statistics are asymptotically distributed as chi-square variates under the null hypothesis, and are asymptotically distribution free (ADF) (Browne, 1982, 1984; Foldnes et al., 2012). The residual based test statistic, T_{RADF}, is given by

$$T_{RADF} = n\hat{e}^T \{\dot{\sigma}_c(\hat{\theta})[\dot{\sigma}_c(\hat{\theta})^T \hat{\Gamma} \dot{\sigma}_c(\hat{\theta})]^{-1} \dot{\sigma}_c(\hat{\theta})^T\} \hat{e} \tag{3.19}$$

Where $\hat{e} = s - \sigma(\hat{\theta})$, $\dot{\sigma}_c(\hat{\theta})$ is a $p^* \times (p^* - q)$ full-rank orthogonal complement of $\dot{\sigma}(\hat{\theta})$, and $\hat{\Gamma}$ is a sample estimate of Γ. For models that are correctly specified, T_{RADF} is asymptotically distributed as a central chi-square variate with d degrees of freedom (Browne, 1984).

Just like T_{ADF}, T_{RADF} typically requires very large samples in order to be correctly distributed (Bentler & Yuan, 1999; Yuan & Bentler, 1998). Thus, Yuan and Bentler (1998) suggested a small sample corrected version:

$$T_{CRADF} = \frac{T_{RADF}}{1 + \frac{NT_{RADF}}{n^2}} \tag{3.20}$$

Where N is the sample size and $n = N-1$. Limited simulation work on this statistic shows that T_{CRADF} may over-correct T_{RADF} at smaller sample sizes (Bentler & Yuan, 1999; Yuan & Bentler, 2003). Neither T_{RADF} nor T_{CRADF} will be defined unless $[\dot{\sigma}_c(\hat{\theta})^T \hat{\Gamma} \dot{\sigma}_c(\hat{\theta})]^{-1}$ in Equation (3.19) is invertible.

3.2 Multilevel factor analysis

The conventional single-level factor analytic procedures described above are all predicated on the assumption that the vectors of individual scores are statistically independent. That is, one individual's scores do not affect another individual's scores. Statistical independence implies that two variables are uncorrelated. When individuals are associated with groups (for example, in the case where students are associated with classrooms) this independence assumption is likely to be violated, and observations within a particular group are likely to be more similar than those across groups, and thus, those observations are likely to be correlated. Thus, analyzing data that is multilevel in structure with conventional factor analysis techniques may raise methodological issues. Specifically, it has been shown that the non-independence of individuals within groups can bias parameter estimates, test statistics, and standard errors (Julian, 2001) when factor analysis is conducted in a conventional framework.

One approach approaches to expanding factor analysis and covariance structure analysis methods to multilevel data (e.g., Goldstein, 2003; Lee, 1990; McDonald & Goldstein, 1989; B. O. Muthén, 1991, 1994) uses a conventional factor analysis

structure on each of the levels, and is built from a multivariate random-intercepts model (i.e., a random-effects MANOVA). For two level data, consider the score decomposition (Lee & Poon, 1998; Longford & Muthén, 1992; Yuan & Bentler, 2007):

$$y_{ij} = \mu + u_j + e_{ij} \tag{3.21}$$

where y_{ij} is a p-variate vector of observed scores for individual i in group j that can be decomposed into a vector of means (μ) and independent between-group (u_j), and within-group (e_{ij}) random components. Because the between-group and within-group random components are independent, the covariance of the observed scores can be expressed as a sum of between-group and within-group covariance components:

$$\Sigma_T = \Sigma_B + \Sigma_W \tag{3.22}$$

where Σ_T, Σ_B and Σ_W are symmetric $p \times p$ covariance matrices. The covariance matrices can be expressed in two separate factor models, one for the between-groups level:

$$\Sigma_B = \Lambda_B \Phi_B \Lambda_B^T + \Psi_B \tag{3.23}$$

and another for the within-group level

$$\Sigma_W = \Lambda_W \Phi_W \Lambda_W^T + \Psi_W \tag{3.24}$$

where Λ_B is a $p \times k$ matrix of factor loadings for p items on k factors, and Λ_W is a $p \times r$ matrix of factor loadings for p items on r factors. Note that while it is possible for $k = r$ and for $\Lambda_B = \Lambda_W$, this is not necessary. Φ_B and Φ_W are $k \times k$ and $r \times r$ matrices of factor covariances, respectively, and Ψ_B and Ψ_W are $p \times p$ diagonal matrices containing unique (residual) variances. It follows that Φ_B need not equal Φ_W, and Ψ_B need not equal Ψ_W.

3.2.1 The segregating approach: multilevel factor analysis using multiple single level models

Because the multilevel factor analysis approach described above uses a conventional factor analysis model on each level, one naturally approach to model testing that emerges is to perform separate, conventional factor analyses on an estimate of each covariance matrix, one at a time. In a two level model, say in a case with students nested within classrooms, this would suggest a two-stage process (Goldstein, 2003; Longford & Muthén, 1992; Yuan & Bentler, 2007). First, estimates of Σ_B and Σ_W are obtained. Then, separate factor analyses on the student level and classroom level covariance matrices are conducted. These matrices can be analyzed "using any standard procedure" (Goldstein, 2003, p. 190). In fact, some of the earliest methodological writings on multilevel factor analysis apply this approach (Cronbach, 1976; Härnqvist, 1978). Ryu and West (2009) refer to this as the "segregating" approach (p. 592).

The first step in implementing the segregating approach is to estimate the within-level and between-level covariance matrices. Two different ways of obtaining estimates are described briefly below. For balanced data, the estimates of these two matrices are unbiased, even when the data is not normally distributed (B. O. Muthén, 1994). For unbalanced data, the estimates are consistent (e.g, Goldstein, 2003; B. O. Muthén, 1994)

3.2.1.1 Obtaining estimated covariance matrices using a multilevel-multivariate model

Goldstein (2003) proposed a multilevel multivariate model to obtain estimates of Σ_B and Σ_W, where the p-vector of observed scores for individual i in group j described above is modeled using a three level hierarchical model. For illustration purposes, consider a case with 6 observed variables. At the first level, there is a

measurement model

$$y_{pij} = \pi_{1ij}d_{1ij} + \pi_{2ij}d_{2ij} + \pi_{3ij}d_{3ij} + \pi_{4ij}d_{4ij} + \pi_{5ij}d_{5ij} + \pi_{6ij}d_{6ij} \tag{3.25}$$

Where $d_{1ij} \cdots d_{6ij}$ are indicator variables that indicate whether the observed score y_{pij} is associated with the 1st, 2nd, 3rd, 4th, 5th or 6th item. There is no residual variation at this level.

At the second level,

$$\pi_{pij} = \beta_{pj} + u_{pij} \tag{3.26}$$

Where the u_{pij}'s represent random effects such that $u_{pij} \sim N(0, \Sigma_W)$. At the third level,

$$\beta_{pj} = \gamma_p + u_{pj} \tag{3.27}$$

Where the u_{pj}'s represent random effects such that $u_pj \sim N(0, \Sigma_B)$. Estimates of Σ_B and Σ_W can then be obtained using REML (restricted maximum likelihood) estimation in a wide variety of multilevel modeling software packages.

3.2.1.2 maximum likelihood estimation

Longford and Muthén (1992) derived estimators of Σ_B and Σ_W based on sums of squares and cross products. Specifically,

$$\hat{\Sigma}_W = \frac{T_1}{N - J} \tag{3.28}$$

and

$$\hat{\Sigma}_B = \frac{T_2 - J(N-J)^{-1}T_1}{N - \frac{\sum_j n_j^2}{N}} \tag{3.29}$$

where

$$T_1 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - y_{.j})(y_{ij} - y_{.j})^T \tag{3.30}$$

and

$$T_2 = \sum_{j=1}^{J} n_j (y_{\cdot j} - y_{\cdot\cdot})(y_{\cdot j} - y_{\cdot\cdot})^T \tag{3.31}$$

where N is the number of individuals, J is the number of groups, n_j is the number of individuals in the jth group, $y_{\cdot j}$ is the mean of the jth group, and $y_{\cdot\cdot}$ is the grand mean.

3.3 Model testing and test statistics in multilevel factor analysis

Because the segregating approach works by applying conventional factor analysis on the separate estimated covariance matrices, model testing can take place in a conventional framework. This dissertation focuses on the analysis of the between-teacher level covariance matrix, since the teacher is often the unit of analysis for the student surveys. In the context of two two-level models, this means that we are focused on the analysis of $\hat{\Sigma}_B$. There are some specific conditions that distinguish model testing on $\hat{\Sigma}_B$ from conventional factor analysis, particularly when ML estimation is used to estimate model parameters and the likelihood ratio test statistic is used to test models on $\hat{\Sigma}_B$. Thus, the multilevel analogs of these conventional test statistics are defined and the specific issues that may arise are discussed below.

3.3.1 ML estimation and the analysis of the between-level covariance matrix

In using the segregating method, estimates of model parameters, $\hat{\theta}$, are found by minimizing the same maximum likelihood discrepancy function defined in Equation 3.12.

$$F_{ML} = log|\Sigma(\theta)| + tr[S\Sigma_B(\theta)^{-1}] - log|S| - p \tag{3.32}$$

In using the segregating method on the between-level covariance matrix, the sample estimate of the covariance matrix S is given by $\hat{\Sigma}_B$. Again, T_{ML} can be defined as:

$$T_{ML} = n\hat{F}_{ML} \tag{3.33}$$

Where $n = J - 1$, one less than the number of groups.

In using the segregating method to analyze $\hat{\Sigma}_B$, T_{ML} is often expected to converge to a central chi-square distribution with d degrees of freedom if the model is correct and there is no excess skew or kurtosis in the observed variables. Several sources (Goldstein, 2003; Hox, 2010; Hox & Maas, 2004; Ryu & West, 2009) suggested that T_{ML} will behave in this way and can be used to evaluate between-level measurement models.

In practice, however, and contrary to the advice given in these sources, T_{ML} may be inflated, and may not have the correct asymptotic distribution, even when the data is normally distributed and the model is correctly specified. This occurs because the sample variability of $\hat{\Sigma}_B$ is larger than what it would have been without the effect of the hierarchical structure, which is commonly called the clustering effect (e.g., Yuan & Bentler, 2006, 2007). The impact of the clustering effect is related to 1) the proportion of total observed variance attributable to group membership (i.e., the ICCs of the observed variables) and 2) within-group sample size.

The intraclass correlation (ICC) represents the proportion of observed variance attributable to group membership, and can be obtained from the diagonal elements of Σ_B and Σ_W. For any given item p, the intraclass correlation for that item can be expressed:

$$ICC_p = \frac{\Sigma_{Bpp}}{\Sigma_{Bpp} + \Sigma_{Wpp}} \tag{3.34}$$

where Σ_{Bpp} and Σ_{Wpp} are diagonal elements of Σ_B and Σ_W respectively. ICC values range between 0 and 1, and for fixed Σ_B, ICCs will increase as the elements of $\Sigma_W \to 0$.

When ICCs are low or within-group sample sizes are small, T_{ML} will not converge in distribution to a centrally distributed chi-square variate, even when the model is correct. As a result, inferences about model structure based on T_{ML} may not be valid for the segregated analysis of $\hat{\Sigma}_B$ even when the data is normally distributed. The clustering effect will also have adverse effects on the estimated standard errors. Even when the model is correctly specified, ML estimation would result in standard errors that were not consistent. This is because, in the segregated analysis of the between-level covariance matrix, ML standard errors are estimated based on

$$Cov\left(\hat{\theta}\right) = \frac{\left[\dot{\sigma}(\hat{\theta})^T W_{BML} \dot{\sigma}(\hat{\theta})\right]^{-1}}{J} \tag{3.35}$$

Where $W_{BML} = .5D_p^T[\Sigma_B(\hat{\theta})^{-1} \otimes \Sigma_B(\hat{\theta})^{-1}]D_p$ and $\Sigma_B(\hat{\theta})$ is the estimated model-implied between-level covariance matrix. However, because of the clustering effect, when ICCs are low or within-group sample sizes are small, W_{BML} is not correctly specified (Schweig, 2014; Yuan & Bentler, 2006, 2007) in the sense that it does not satisfy Equation 3.5. Based on results presented elsewhere (e.g., Browne, 1984; Hox, 2010; Yuan & Bentler, 2007), it is anticipated that the parameter estimates themselves will be consistent.

This fact is rarely made explicit in methodological literature on multilevel factor analysis. Even when the poor empirical performance of T_{ML}, or of ML estimates of standard errors is noted (Hox, 2010; Yuan & Bentler, 2007), the possible role of either item ICC or within-group sample are not described. In fact, several sources (Goldstein, 2003; Hox, 2010; Hox & Maas, 2004; Ryu & West, 2009) suggested that the segregating method is a "viable method" (Hox & Maas, 2004, p. 145) that can be "implemented within the preexisting ML . . . framework." (Ryu & West, 2009, p. 600).

3.3.2 Asymptotically Distribution Free methods for factor analysis with segregated matrices

Yuan and Bentler (2003) proposed both ADF test statistics, and residual-based ADF test statistics equivalent to those developed by Browne (1974, 1982, 1984), that are appropriate for multilevel factor analysis. Analogously to the conventional, single level statistic, In the segregated analysis of the between-level covariance matrix the ADF test statistic T_{ADF} is defined as:

$$T_{ADF} = n\left(s - \sigma_B(\hat{\theta})\right)^T \hat{\Gamma}_B^{-1}\left(s - \sigma_B(\hat{\theta})\right) \tag{3.36}$$

In using the segregating method on the between-level covariance matrix, $n = J - 1$ and $s = vech(\hat{\Sigma}_B)$.$\hat{\Gamma}_B$ is a consistent estimate of Γ_B, given by the asymptotic covariance matrix of the between-level covariance matrix:

$$\sqrt{J}(s_B - \sigma_B(\theta)) \xrightarrow{d} N(0, \Gamma_B) \tag{3.37}$$

As developed by Yuan and Bentler (2003, 2007), $\hat{\Gamma}_B^{-1}$ is obtained using generalized estimating equations (GEEs) (e.g., K.-Y. Liang & Zeger, 1986). Essentially, K.-Y. Liang and Zeger (1986) applied the Huber-White sandwich estimator (Huber, 1967; White, 1980) to obtain estimates of the variability of parameter estimates in hierarchically structured data. Yuan and Bentler (2002, 2006, 2007) applied this principle to obtain estimates of the variability of $\hat{\Sigma}_W$ and $\hat{\Sigma}_B$. This estimate of Γ_B^{-1} is called $\hat{\Gamma}_{GEE}^{-1}$ throughout this dissertation. Likewise, the associated estimate of Γ_B, is called $\hat{\Gamma}_{GEE}$.

In addition to T_{ADF}, the corrected test statistic T_{CADF} can also be used in conjunction with the segregating approach. That statistic is given by:

$$T_{CADF} = \frac{T_{ADF}}{1 + \frac{T_{ADF}}{n}} \tag{3.38}$$

where $n = J - 1$.

ADF standard errors can be obtained similarly for the segregating method as in conventional, single level models. Standard errors (Yuan & Bentler, 2006) are given by

$$Cov\left(\hat{\theta}\right) = \frac{\left[\dot{\sigma}_B(\hat{\theta})^T \hat{\Gamma}_B^{-1} \dot{\sigma}_B(\hat{\theta})\right]^{-1}}{J} \tag{3.39}$$

Given sufficient sample sizes and a correctly specified model, T_{ADF} and T_{CADF} will be correctly distributed under the null hypothesis, and the standard errors will be consistent. This is because the generalized estimating equations used to estimate $\hat{\Gamma}_{GEE}$ take into account the excess sampling variability in $\hat{\Sigma}_B$ that results from the clustering (Schweig, 2014). Thus, it is anticipated that T_{ADF} and T_{CADF} will converge to the correct chi-square distribution, regardless of the item ICCs or the number of individuals in each group.

3.3.3 Robust methods for factor analysis with segregated matrices

The multilevel version of the Satorra-Bentler (1988) rescaled test statistic T_{RML}, developed by Yuan and Bentler (2003, 2007) is given by:

$$T_{RML} = \frac{T_{ML}}{k} \tag{3.40}$$

Where $k = \frac{tr(\hat{U}\hat{\Gamma}_B)}{d}$ and $\hat{U}$ is as defined as

$$\hat{U} = W_{BML} - W_{BML}\dot{\sigma}_B(\hat{\theta}) \left(\dot{\sigma}_B(\hat{\theta})^T W_{BML} \dot{\sigma}_B(\hat{\theta})\right)^{-1} \dot{\sigma}_B(\hat{\theta})^T W_{BML} \tag{3.41}$$

Robust standard errors (e.g., Bentler & Dijkstra, 1985; Shapiro, 1986) are based on a triple-sandwich estimator:

$$Cov\left(\hat{\theta}\right) = \frac{A\left[\dot{\sigma}_B(\hat{\theta})^T W_{BML} \hat{\Gamma}_B W_{BML} \dot{\sigma}_B(\hat{\theta})\right] A}{J} \tag{3.42}$$

where J is the number of groups and $A = \left[\dot{\sigma}_B(\hat{\theta})^T W_{BML} \dot{\sigma}_B(\hat{\theta})\right]^{-1}$.
T_{RML} is expected to converge to a distribution with the correct first moment regardless of ICC and within-group sample size. The scaling constant, k will be greater than 1. Bentler (2006) explained that $tr(\hat{U}\hat{\Gamma}_B)$can be thought of as a way to determine the discrepancy between the hypothesized model and data distribution (carried by $\hat{U}$) and the true data distribution (carried by $\hat{\Gamma}_B$). In analyzing $\hat{\Sigma}_B$, the discrepancy between $\hat{U}$ and $\hat{\Gamma}_B$ occurs because $\hat{\Gamma}_B$ accounts for the clustering effect, and $\hat{U}$ does not (Schweig, 2014).

3.3.4 Residual-based test statistics with segregated matrices

The residual-based test statistics for use in the segregating approach were presented by Yuan and Bentler (2007). The residual-best test statistic T_{RADF} is given by

$$T_{RADF} = n\hat{e}^T\{\dot{\sigma}_{Bc}(\hat{\theta})[\dot{\sigma}_{Bc}(\hat{\theta})^T\hat{\Gamma}\dot{\sigma}_{Bc}(\hat{\theta})]^{-1}\dot{\sigma}_{Bc}(\hat{\theta})^T\}\hat{e} \tag{3.43}$$

Where $\hat{e} = s_B - \sigma_B(\hat{\theta})$, and $n = J - 1$. Yuan and Bentler (2007), the expression for the corrected test statistic T_{CRADF} is slightly different from the conventional, single level version given in Equation 3.20.

$$T_{CRADF} = \frac{T_{RADF}}{1 + \frac{T_{RADF}}{J}} \tag{3.44}$$

the difference between the correction given by Equation 3.44 and Equation 3.20 will be minimal for sufficient sample sizes (e.g., Bentler, 2006).

3.4 Application of conventional factor analysis to multi-level data

The segregating approach described above provides a method to apply factor analytic techniques to data that is hierarchical in structure. However, obtaining estimates of within-level and between-level covariance matrices is considerably more involved than obtaining sample covariance matrices in conventional factor analysis. Partly because of these technical complications, and partly because, as Sirotnik (1980) and Marsh et al. (2012) have noted, researchers have not always been careful in their consideration of unit of analysis issues during instrument development, it is common to find examples in the research literature where conventional factor analysis procedures are often applied either on the disaggregated data, or on the group means (weighted or unweighted). The problems that can arise from these approaches were described briefly in Chapter 2. In particular, using conventional factor analysis imposes strict measurement invariance constraints across the within-level and between-level models, and ignoring clustering can result in biased parameters and standard errors and inflated test statistics (e.g. Julian, 2001; Preacher et al., 2010). Here, the statistical foundations of these problems are described in more detail.

3.4.1 Factor analysis of the disaggregated covariance matirx

In this case, a confirmatory factor analysis is performed on an estimate of Σ_T in Equation 3.22 without attributing variance to distinct within and between sources. This approach imposes cross-level measurement invariance constraints on the model. That is, this approach assumes, *de facto*, that $\Lambda_B = \Lambda_W$, $\Phi_B = \Phi_W$, and $\Psi_B = \Psi_W$ as given in Equations 3.24 and 3.23 (e.g., Zyphur et al., 2008). There is a long history of methodological work (e.g. Cronbach, 1976; Härnqvist, 1978; Longford & Muthén, 1992; Zyphur et al., 2008) that shows that this assumption may not be met

in practice. Moreover, as Zyphur et al. (2008) noted, "discovering differential factor structures across levels of analysis may be of substantive interest to researchers." (p. 130), in the sense that researchers may be interested in developing separate theories for within-level and between-level phenomena (e.g., Cronbach, 1976), and that investigating differences in the networks of correlates (e.g., Guion, 1973) of individual and group level variables may be of substantive importance.

A simulation study by Julian (2001) demonstrated that, even when the number of factors at the within-level and between-level were the same, using conventional factor analysis procedures in this way can lead to biased parameters and standard errors, and inflated test statistics (Julian, 2001; Preacher et al., 2010, see also). The problems of test-statistic inflation and parameter bias increased as the intraclass correlation of the items increased—the more variance in observed indicators that is attributed to groups, the greater the extent of parameter bias, standard error bias, and test statistic inflation. A simulation study by J.-Y. Wu and Kwok (2012) showed similar results in terms of standard errors and parameter estimates. J.-Y. Wu and Kwok (2012) also found that the model test statistics obtained from a disaggregated factor analysis were not helpful for testing the between-level factor structure. As they noted, "overall model chi-square test and commonly used fit indexes could not consistently provide much helpful information on the necessity of specifying a different higher level model" (p. 31).

This trend parallels research in univariate multilevel models (Raudenbush & Bryk, 2002), which shows that ignoring the hierarchical structure of a data set can bias standard errors. Overall, they suggest that analyzing multilevel data with conventional factor analytic techniques can lead to "substantively misleading" inferences about relationships among indicators, or between indicators and external variables (Reise et al., 2005, p. 130).

3.4.2 Group-means factor analysis

In this case, factor analysis is performed on either the unweighted sample covariance matrix of group-means

$$S_B^* = \frac{1}{J}\sum_{j=1}^{J}(y_{\cdot j} - y_{\cdot\cdot})(y_{\cdot j} - y_{\cdot\cdot})^T \tag{3.45}$$

or the sample covariance matrix of group-means weighted by group size:

$$S_B = \frac{1}{J}\sum_{j=1}^{J} n_j(y_{\cdot j} - y_{\cdot\cdot})(y_{\cdot j} - y_{\cdot\cdot})^T \tag{3.46}$$

The problem with using the group mean covariance matrix as given by Equation 3.45 or Equation 3.46 is that these sample matrices are not consistent estimators of the population between-level covariance matrix Σ_B. For the sake of illustrative clarity, consider a situation where each group has the same sample size n (i.e., the case of balanced groups). Results in B. O. Muthén (1994) showed that S_B in Equation 3.46 is a consistent and unbiased estimator of

$$\Sigma_W + n\Sigma_B \tag{3.47}$$

That is, S_B as given in Equation 3.46 contains both within and between variance sources. Thus, B. O. Muthén (1994) noted that "any simple structure expected to hold for Σ_B does not necessarily hold for S_B" (p. 388). The unweighted covariance matrix given by Equation 3.45 presents additional complications, by giving "small groups and large groups equal weight in determining parameter estimates." (Preacher et al., 2010, p. 213). Ignoring differences in group size may introduce additional biases into model testing.

3.5 A note about other approaches to multilevel factor analysis

The segregating approach that is adopted and expanded in this dissertation is only one potential approach to multilevel factor analysis. Another approach is based on a multilevel ML discrepancy function (e.g., Lee, 1990; Lee & Poon, 1992; B. O. Muthén, 1991) that finds optimal model parameters simultaneously for both the within-level and between-level covariance structure models. A two-level ML discrepancy function, can be expressed as:

$$F_{ML} = \sum_{j=1}^{J}(n_j - 1)\{log|\Sigma_W(\theta)| + tr(\Sigma_W(\theta)^{-1}S_{yWj})\} + \sum_{j=1}^{J}\{log|\Sigma_W(\theta) + \frac{1}{n_j}\Sigma_B(\theta)| + tr([\Sigma_W(\theta) + \frac{1}{n_j}\Sigma_B(\theta)]^{-}1S_{gj}) \quad (3.48)$$

as given in Ryu and West (2009, p. 586). N is the number of individuals, J is the number of groups, and n_j is the number of individuals in the jth group. $S_{yWj} = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - y_{\cdot j})(y_{ij} - y_{\cdot j})^T$ and $S_{gj} = n_j(y_{\cdot j} - y_{\cdot\cdot})(y_{\cdot j} - y_{\cdot\cdot})^T$. Optimal parameter estimates, $\hat{\theta}$, are found by minimizing the discrepancy function given in Equation 3.48 (Lee, 1990).

This discrepancy function yields two potential approaches to test measurement models. Models can simultaneously be specified on the within-level and between-level structures (B. O. Muthén, 1994; Ryu & West, 2009; Yuan & Bentler, 2007; Hox, 2010), in what can be called the "simultaneous approach", or an unrestricted (saturated) model can be fit at one level, and a measurement model can be tested on the other level (Hox, 2010; B. O. Muthén, 1994; Ryu & West, 2009), referred to as the "partially saturated model method" (Ryu & West, 2009, p. 589).

In the simultaneous approach, a likelihood ratio test statistic can be obtained

(Bentler & Liang, 2003) from Equation 3.48 as:

$$T_{ML} = F_{ML}[\hat{\theta}_W, \hat{\theta}_B] - F_{ML}[\hat{\theta}_{WS}, \hat{\theta}_{BS}] \tag{3.49}$$

Where $\hat{\theta}_{WS}$ refers to the set of parameters that structure the within-level covariance matrix, and $\hat{\theta}_{BS}$ refers to the set of parameters that structure the between-level covariance matrix. This test statistic can be used to test the null hypothesis that the population covariance matrix, $\Sigma_T = \Sigma_B + \Sigma_W$, is equal to the model implied covariance matrix, $\Sigma_T(\theta) = \Sigma_B(\theta) + \Sigma_W(\theta)$. In the partially saturated model method, an unrestricted model is specified at one level. If an unrestricted model is fit to the within-level, 3.49 becomes:

$$T_{ML} = F_{ML}[\hat{\theta}_{WS}, \hat{\theta}_B] - F_{ML}[\hat{\theta}_{WS}, \hat{\theta}_{BS}] \tag{3.50}$$

Thus, any lack of fit in the model is attributed to the discrepancy between $\Sigma_B(\theta)$ and $\hat{\Sigma}_B$. (e.g., Ryu & West, 2009). In other words, since the saturated model estimates all covariances between variables, it has no degrees of freedom, and the within-level model makes no contribution to the chi-square test statistic.

While the simultaneous approach is widely used in the applied literature (e.g., Dyer, Hanges, & Hall, 2005; Holfve-Sabel & Gustafsson, 2005; Sexton et al., 2006) it has been shown in several studies (e.g., Hox, 2010; Ryu & West, 2009; Yuan & Bentler, 2007) that simultaneously modeling the within-level and between-level structures does not produce meaningful diagnostic information about the between-level factor structure. Thus, the simultaneous modeling of between and within factor structures makes model or theory revision difficult (Yuan & Bentler, 2007), and this approach is not recommended in the literature. The partially saturated model method, on the other hand, does provide level-specific diagnostic information, because the chi square test statistic reflects fit (or misfit) on only one level, but this approach was not meant to provide parameter estimates or standard errors (Ryu & West, 2009;

Yuan & Bentler, 2007). A practical issue with this method is that estimates of fit indices such as the Root Mean Square Error of Approximation (RMSEA) (Steiger & Lind, 1980), and the Comparative Fit Index (CFI) (Bentler, 1990) provided by software programs will spuriously show good fit (Hox, 2010, p. 307), and so may be misinterpreted (e.g., Kunter et al., 2008; Rosenberg, 2009).

3.6 Extensions to three levels of nesting: students, teachers, schools.

Thus far, the segregating approach has only been described in the context of two level models, for example, in the case where students are nested within teachers. This section outlines how this approach can be expanded to three level models, for example, in the case where students are nested within teachers, nested within schools. First, we may consider extending the score decomposition (Longford & Muthén, 1992) given in 3.21:

$$y_{ijk} = \mu + v_k + u_{jk} + e_{ijk} \tag{3.51}$$

where the vector of observed scores for individual i in subgroup j at the second level in group k at the third level (y_{ijk}) can be decomposed into independent level-3 (v_k) , level-2 (u_{jk}) and level-1 (e_{ijk}) random components. Using the notation of Yau, Lee, and Poon (1993), the covariance of the observed scores can be expressed as a sum:

$$\Sigma_T = \Sigma_B + \Sigma_{WG} + \Sigma_{WGK} \tag{3.52}$$

Where Σ_B, Σ_{WG} and Σ_{WGK} are symmetric positive definite $p \times p$ covariance matrices. These covariance matrices can be expressed in separate factor models, analogously to the two-level models given in Equation 3.23 and Equation 3.24. While Yau et al. (1993) describe a general class of covariance structures where the

level-2 and level-1 covariance structures may differ across groups, this dissertation considers a particular class of three level covariance structure models where it is assumed that the same covariance structure holds across subgroups and across individuals within subgroups. This class of models is a specific case of the more general structures described by Yau et al. (1993).

For example, in the case of student ratings of instructional practice, where students are nested within teachers nested within schools, the structure of Σ_{WG}, may be of particular interest, as this describes the teacher-level covariance structure. Thus, a set of parameters may be selected such that

$$\Sigma_{WG} = \Lambda_{WG}\Phi_{WG}\Lambda_{WG}^T + \Psi_{WG} \tag{3.53}$$

As in the case of the two level models described earlier, a consistent estimate of Σ_{WG} can be found using the multilevel-multivariate models described by Goldstein, or by using sums and crossproducts of squares (e.g., Goldstein, 2003; Longford & Muthén, 1992). Consistent estimates can also be obtained using commercial software products, such as Mplus (L. K. Muthén & Muthén, 2010).

Longford and Muthén (1992) suggested that a consistent estimate of Σ_{WG} could be found based on withinsubgroup and subgroup-within-group sums of squares:

$$\hat{\Sigma}_{WG} = \frac{T_2 - (J-K)(N-J)^{-1}T_1}{N - \sum_k \frac{\sum_j n_{jk}^2}{N_k}} \tag{3.54}$$

Where

$$T_1 = \sum_{k=1}^{K}\sum_{j=1}^{n_k}\sum_{i=1}^{n_{jk}}(y_{ijk} - y_{\cdot jk})(y_{ijk} - y_{\cdot jk})^T \tag{3.55}$$

and

$$T_2 = \sum_{k=1}^{K}\sum_{j=1}^{n_k} n_{jk}(y_{\cdot jk} - y_{\cdot\cdot k})(y_{\cdot jk} - y_{\cdot\cdot k})^T \tag{3.56}$$

And N is the number of individuals, J is the number of subgroups, and K is the number of groups. n_{jk} is the number of individuals in the jth subgroup in the kth

group, n_k is the number of subgroups in the kth group, and $N_k = \sum_j n_{jk}$. $y_{\cdot jk}$ is the vector of means for the jth subgroup of the kth group, and $y_{\cdot\cdot k}$ is the vector of means for the kth group.

Recall that in conventional factor analysis, the asymptotic covariance matrix Γ is given by the fact that $\sqrt{n}(s - \sigma(\theta)) \xrightarrow{d} N(0, \Gamma)$, and analogously, in the segregated analysis of Σ_B, $\sqrt{J}(\hat{\sigma}_B - \sigma(\theta)_B) \xrightarrow{d} N(0, \Gamma_B)$. Results in Yuan and Bentler (2006) and Yau et al. (1993, p. 170) imply that this also holds true for $\sigma_{WG} = vech(\Sigma_{WG})$:

$$\sqrt{J-K}(\hat{\sigma}_{WG} - \sigma(\theta)_{WG}) \xrightarrow{d} N(0, \Gamma_{WG}) \tag{3.57}$$

As such, the six test statistics (T_{ADF}, T_{CADF}, T_{ML}, T_{RML}, T_{RADF} and T_{CRADF}) presented in a two level framework can readily be implemented in the analysis of $\hat{\Sigma}_{WG}$ by using $\hat{\Gamma}_{WG}$, a consistent estimate of Γ_{WG} as the asymptotic covariance matrix, and specifying $J-K$ as the sample size. T_{ML} is still anticipated to perform poorly, because of the clustering effect. However, T_{ADF}, T_{CADF}, T_{RML}, T_{RADF} and T_{CRADF} use $\hat{\Gamma}_{WG}$ and are anticipated to perform well for sufficiently large sample sizes.

3.7 Multilevel models with level-restricted variation

The models in Equation 3.21 and Equation 3.51 do not include level-2 or level-3 observations. That is, those models do not include variables that are observed at either the subgroup or the group levels. In school or classroom research, it may be that there are variables describing classroom level phenomenon that are measured at the classroom level. These variables may include, for example, information about attendance rates or available economic resources. J. Liang and Bentler (2004) generalized the formulations given in Equation 3.21, which were presented in Lee and Poon (1998). In this two-level formulation, y_{ij} refers to variables observed at the individual level, and z_j to variables that are measured at the group level.

Given:

$$\begin{pmatrix} z_j \\ y_{ij} \end{pmatrix} = \begin{pmatrix} z_j \\ u_j \end{pmatrix} + \begin{pmatrix} 0 \\ e_{ij} \end{pmatrix} \tag{3.58}$$

where $\mu = E \begin{pmatrix} z_j \\ u_j \end{pmatrix} = \begin{pmatrix} \mu_z \\ \mu_u \end{pmatrix}$, $\Sigma_B = Cov \begin{pmatrix} z_j \\ u_j \end{pmatrix} = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zu} \\ \Sigma_{uz} & \Sigma_{uu} \end{pmatrix}$, and $\Sigma_W = Cov(e_{ij})$. An estimate of Σ_B can be obtained using Equation 3.29.

This formulation can also be extended to three levels as:

$$\begin{pmatrix} q_k \\ z_{jk} \\ y_{ijk} \end{pmatrix} = \begin{pmatrix} q_k \\ w_k \\ v_k \end{pmatrix} + \begin{pmatrix} 0 \\ s_{jk} \\ u_{jk} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ e_{ijk} \end{pmatrix} \tag{3.59}$$

where $\mu = E \begin{pmatrix} q_k \\ w_k \\ v_k \end{pmatrix} = \begin{pmatrix} \mu_q \\ \mu_w \\ \mu_v \end{pmatrix}$, $\Sigma_B = Cov \begin{pmatrix} q_k \\ w_k \\ v_k \end{pmatrix} = \begin{pmatrix} \Sigma_{qq} & \Sigma_{qw} & \Sigma_{qv} \\ \Sigma_{wq} & \Sigma_{ww} & \Sigma_{wv} \\ \Sigma_{vq} & \Sigma_{vw} & \Sigma_{vv} \end{pmatrix}$, $\Sigma_{WG} = Cov \begin{pmatrix} s_{jk} \\ u_{jk} \end{pmatrix} = \begin{pmatrix} \Sigma_{ss} & \Sigma_{su} \\ \Sigma_{us} & \Sigma_{uu} \end{pmatrix}$, and $\Sigma_{WGK} = Cov(e_{ijk})$ An estimate of Σ_{WG} can be obtained using Equation 3.54.

3.8 Open issues in the literature on multilevel factor analysis

A growing research base points to significant problems when applying conventional factor analysis to multilevel data, or simultaneously modeling the between and within factor structures. At the same time, number of additional issues in multilevel factor analysis are not addressed in the existing literature. First, the existing work does not address how common test statistics and estimation methods are likely to perform under conditions likely encountered in real-world settings and contexts. Second, there are only a handful of studies that empirically investigate differences

in test statistic performance and estimation efficiency across the saturating and segregating approaches, and how these differences should influence modeling strategies (Hox & Maas, 2004; Ryu & West, 2009). Third, existing work does not address the performance of GEE methods to obtain $\hat{\Gamma}_{GEE}$ under conditions likely encountered in real-world settings. Fourth, existing literature does not offer empirical demonstrations of the segregating approach with additional levels of nesting (hierarchical data sets with three or four levels, for example). Each of these issues is addressed in turn in the sections that follow.

3.8.1 Relative efficiency across the saturating and segregating approaches

Parameter estimates obtained from two-stage approaches like the segregating approach are, in general, less efficient than those obtained from one-stage approaches like the partially-saturated model method described above. Yuan and Bentler (2007) suggested that the simultaneous estimators are asymptotically most efficient, and so there may be a gain in efficiency in using the partially saturated model method over the segregating approach. Yuan and Bentler (2007) demonstrated that, as either the within-group sample size or the number of groups get larger, the difference in efficiency between these two approaches will be small. Yuan and Bentler (2007) did suggest that, in small to medium sized samples, particularly with larger models, the segregating approach may actually be more efficient than either of these approaches, because parameter estimates based on a smaller model (the segregating approach will, in general, have far fewer parameters than the other two approaches) will have more numerical stability (Yuan & Bentler, 2007, p. 56). Longford and Muthén (1992) noted that the loss of efficiency in this procedure "may be quite modest" (p. 589).

While some studies examine parameter bias in both the partially saturated model method and the segregating approach (e.g., Hox & Maas, 2004; Hox, 2010), thus far, the author is unaware of any systematic exploration of the comparative efficiency

of these approaches in the literature.

3.8.2 Test statistic performance and parameter estimation under real world conditions

Methodological work on multilevel factor analysis often contains simulation conditions not likely to be encountered in real settings with student surveys. For example, Ryu and West (2009) noted that they only included a high ICC condition in their study ($ICC = 0.5$), so that their results may not generalize to other conditions. Furthermore, their simulated model is relatively simple (only 8 degrees of freedom in the between-level model), and considers only within-group sample sizes of 30, 50 and 100. Yuan and Bentler (2007) used only one ICC condition (with an average item ICCs of approximately .37), and the within-group sample sizes are uniformly distributed on [6, 205], which yields an average group size of approximately 106 individuals. The model is also relatively simple (19 degrees of freedom). Other simulation studies (e.g., Hox & Maas, 2004; Hox et al., 2010) also use relatively small models and large sample sizes.

These simulation studies serve an important purpose in that they offer an "empirical verification" (Yung & Bentler, 1994, p. 66) of the model testing methods outlined in previous sections. However, as Yung and Bentler (1994) point out, "these results do not provide evidence to advocate the use [of a particular test statistic] because researchers seldom have such a small model with a large sample size" (p. 66). Indeed, these simulation conditions do not represent those likely encountered in analyses of student surveys. For example, Toland and De Ayala (2005) reported a study with 54 classrooms, average within-class sample sizes of 15, and item ICCs ranging from .06 to .43, with between-factor models with over 150 degrees of freedom. Holfve-Sabel and Gustafsson (2005) had 60 classrooms, average within-group sample sizes of approximately 25 student, and a between-level model with over 500 degrees of freedom in their study. Marsh et al. (2012) noted

that, generally speaking, item ICCs for climate variables are "often less than .1 and rarely greater than .3" (p. 115). Under these conditions, the performance of both the segregating approach and the partially saturated model method, both in terms of parameter bias and variability, and the performance of many commonly used test statistics is largely unknown.

3.8.3 Estimation of the asymptotic covariance matrix

One particular concern in implementing the segregating approach is that obtaining good estimates of test statistics and standard errors that can be used to make valid inferences about measurement models relies heavily on the quality of the estimation of the asymptotic covariance matrix (Γ_B in Equation 3.6) (Hu et al., 1992; Yung & Bentler, 1994). Even with no excess skew or kurtosis in the distribution of the observed variables, the likelihood ratio test statistic T_{ML} will not be correctly distributed under the null hypothesis and the maximum likelihood estimates of standard errors are expected to be biased when analyzing the between-group covariance matrix, particularly when ICCs are low and group sizes are small (Schweig, 2014). Because of this, research has recommended that the segregating approach should be used in conjunction with ADF, residual based, or rescaled test statistics (Schweig, 2014; Yuan & Bentler, 2007), which require an estimate of Γ_B in order to be computed.

The relationship between the estimation of Γ and the performance of ADF estimators has been demonstrated in the context of single level conventional factor analysis, where numerous studies have shown that ADF standard errors tend to be downwardly biased, and test statistics over-reject correct models at all but the largest sample sizes (e.g., Curran et al., 1996; Hox et al., 2010; Hu et al., 1992; B. O. Muthén & Kaplan, 1985, 1992; Powell & Schafer, 2001). In the context of conventional factor anlaysis, Hu et al. (1992) suggested that the poor performance of the ADF method was related to the poor estimation of the elements of Γ, and

empirical results in Yung and Bentler (1994) supported this claim.

However, while there is a relatively robust tradition of literature investigating the estimation of Γ in conventional factor analysis, this author has found no studies that investigate the accuracy with which Γ_B is estimated in the case of multilevel data. In the single level conventional factor analysis framework, the asymptotic covariance matrix is typically estimated directly from the fourth-order sample moments of the observed data. In the segregating approach, the estimation is more numerically complex. Specifically, in the segregating approach, the estimation of Γ_B is based on a sample covariance matrix (Σ_B) that is, itself, the result of a complex estimation process. This may suggest that the estimation of Γ_B is even less accurate for use with the segregating method than it is in conventional single level analyses.

Yuan and Bentler (2002, 2006, 2007) applied the principle of Generalized Estimating Equations (GEE) to estimate Γ_B, the asymptotic covariance matrix for $\hat{\Sigma}_B$. Essentially, this entails an application the Huber-White sandwich estimator (Huber, 1967; White, 1980) to obtain estimates of the variability of the covariance matrix. An alternative to using a Generalized Estimating Equation (GEE) approach is to use a non-parametric bootstrap-based approach (Davison & Hinkley, 1997; Efron & Tibshirani, 1993). The non-parametric bootstrap proceeds by treating the empirical distribution as a population distribution, and repeatedly resampling from the empirical distribution. This can then be used to obtain information about population parameters. There is research (e.g., Feng, McLerran, & Grizzle, 1996; Mancl & DeRouen, 2001; Sherman & Cessie, 1997) suggesting that non-parametric bootstrap methods outperform GEE methods in certain conditions, and that the variance components obtained based on generalized estimating equations can be biased, particularly if the sample size is small (Sherman & Cessie, 1997, p. 908). However, the non-parametric bootstrap has not been applied to the estimation of Γ_B, and there is no information about the comparative performance of the GEE

and cluster-bootstrap approaches in this context.

3.8.4 Multiple levels of nesting: looking beyond two-level models

The vast majority of writing about multilevel factor analysis considers only explicitly cases where there are two levels of nesting in the hierarchical structure. Persons nested within groups. There is little guidance on how to handle additional levels of hierarchy, and little guidance on how ignoring these additional levels would influence model test statistics and parameter estimates. For example, in school settings, it would be common to encounter hierarchical data structures with three levels: students nested in teachers nested in schools. In fact, in secondary school settings, where individual teachers may teach several sections of the same class, it is possible to conceive of hierarchical data structures with four levels: students nested in classes, classes nested in teachers, and teachers nested in schools. While several key papers suggest that extensions to multiple levels is conceptually straightforward (Longford & Muthén, 1992; Yuan & Bentler, 2007), most of the examples in the literature deal with two levels of nesting only. Thus far, the author is unaware of any published research studies on classroom climate deal with this issue explicitly. Holfve-Sabel and Gustafsson (2005) worked with 60 6th grade classrooms, but do not mention the nesting of classrooms into schools, or how this additional level of nesting may influence their findings. Similarly, Van Horn (2003) and D'haenens et al. (2012) look at teacher ratings of school climate using a two-level confirmatory factor analysis, but do not consider the possibility that the nesting of teachers within academic units or departments may influence results.

3.9 Research questions

This dissertation involves a series of analyses of real and simulated student survey data , conducted in order to address a number of the open issues with multilevel

factor analysis described above. Specifically, this dissertation investigates:

1. The efficiency of the segregating approach compared to the partially saturated model method in the estimation of parameters in two-level models.
   (a) How does the efficiency of the segregating approach compare to the partially saturated model method approach for the estimation of between-group factor models?
2. Comparative performance of GEE-based ADF, cluster-bootstrap-based ADF and ML estimators in the segregated analysis of $\hat{\Sigma}_B$
   (a) Are the parameter estimates unbiased using the segregating method?
   (b) How variable are the parameter estimates?
   (c) Are the standard error estimates consistent using the segregating method?
   (d) Are test statistics appropriately distributed under conditions likely to be encountered in student survey research?
   (e) Does the cluster-bootstrap provide a consistent and accurate estimate of the asymptotic covariance matrix?
3. Extension of the bootstrap-based method to three level models.
   (a) Can a bootstrap-based approach to estimating the asymptotic covariance matrix be extended to data sets with three levels?
   (b) Are the parameter estimates unbiased and precise?
   (c) Are test statistics appropriately distributed?
4. The application of these bootstrap methods to a realistic dataset to investigate the dimensions of professional practice that are discernible in a state-wide student survey of instructional quality.

(a) What dimensions of instructional practice are discernible based on student responses in an opportunity to learn survey?

(b) How do these survey-derived variables relate to outcomes of policy interest, such as teacher contributions to student achievement (i.e., teacher value added scores)?

CHAPTER 4

Cluster Bootstrap in Multilevel Factor Analysis

The bootstrap (e.g., Davison & Hinkley, 1997; Efron & Tibshirani, 1993) proceeds by treating the empirical distribution as a population distribution, and repeatedly resampling from the empirical distribution. This can then be used to obtain information about population parameters. Bootstrap methods have been applied in conventional factor analysis (Beran & Srivastava, 1985; Bollen & Stine, 1992), predominantly to obtain information about test statistic distributions. Yung and Bentler (1994) applied the bootstrap to obtain bias-corrected versions of ADF test statistics, and found that these bootstrap corrected statistics performed far better than the uncorrected statistics with sufficiently large sample sizes. Yuan and Hayashi (2006) compared bootstrap and asymptotic standard errors in factor analysis models under a variety of conditions (misspecified models, excessively kurtotic observed variables, etc.), and found that, given finite fourth order moments, even when the model was incorrectly specified, a non-parametric bootstrap can provide consistent estimates of standard errors.

It has been shown that bootstrap based approaches may outperform GEE methods in the estimation of means and variance components in two-level models with small sample sizes (Feng et al., 1996; Sherman & Cessie, 1997). However, many of these investigations consider only univariate models (Sherman & Cessie, 1997), and do not explicitly address the estimation of the asymptotic variances (i.e., the variance of the variance components). To the author's knowledge, no studies comparing the bootstrap and GEE methods for multilevel multivariate models exist.

There is another possible advantage, in addition to the possible improved recovery of the asymptotic covariance matrix, to using a bootstrap-based approach instead of a GEE approach. While GEE approaches are, in theory, extendable to more than two levels of nesting, such an extension requires the researcher to articulate a likelihood function and a vector of first-order partial derivatives, which may be complicated as the number of levels of nesting increases (Yau et al., 1993; Yuan & Bentler, 2007). In contrast, the bootstrap-based approaches are, in principle, readily extendable to multiple levels of nesting and require no explicit articulation of a likelihood function. That said, there is little theoretical or empirical work on the performance of bootstrap methods with nested or clustered data, particularly when there are multiple levels of nesting.

There are many possible approaches to bootstrapping multilevel data. Van der Leeden, Busing, and Meijer (1997) referred to these methods in three different classes: 1) Parametric bootstrap approaches, which use the parametrically estimated distribution function of the data to generate bootstrap samples. (p. 9). 2) Residual bootstrap methods, such as those discussed in Carpenter, Goldstein, and Rasbash (1999), where level-1 and level-2 residuals are resampled, and 3) the non-parametric "cases" bootstrap, where the observed data is resampled with replacement.

4.1 The non-parametric bootstrap for clustered data

This dissertation uses a multilevel version of the non-parametric cases bootstrap, called the cluster bootstrap, where intact clusters (groups) are resampled (Davison & Hinkley, 1997). In a two level model, this would mean resampling at level-2 only. This cluster bootstrap has been shown to provide consistent estimates of variance and covariance components in balanced univariate two-level models (Field & Welsh, 2007), and functions well even for unbalanced univariate data (Samanta

& Welsh, 2013). Recent work by Ren et al. (2010) examined several different variations of the cluster bootstrap and concluded that, for arbitrary group sizes, the best bootstrap scheme samples with replacement from the highest level. Ren et al. (2010) investigated the cluster bootstrap in both two and three level models. In addition, the cluster bootstrap is non-parametric, meaning no distributional assumptions are made on the observed variables. Non-parametric bootstraps such as the cluster bootstrap provide better estimates of model parameters than parametric bootstrap methods under arbitrary distributions, and when there is excessive kurtosis or skew in the observed variables. Results in Samanta and Welsh (2013) suggested that the cluster bootstrap also outperforms other non-parametric methods, such as the residual-based bootstrap.

While some sources (Timmerman, Kiers, Smilde, Ceulemans, & Stouten, 2009; Van der Leeden et al., 1997) have suggested that a multilevel bootstrap resampling plan should reflect which levels in the hierarchy are considered random, (Timmerman et al., 2009, p. 299), and thus, a two stage bootstrap should be used where first groups are resampled, and then individuals are resampled within those groups, theoretical work by Field and Welsh (2007) suggested that this sort of cluster bootstrap does not produce consistent estimators for variance components unless both the number of individuals in each group, n_j, and the number of groups, J, both tend to ∞.

Thus, the non-parametric cluster bootstrap proposed here proceeds in the following way for the two level models (students nested in teachers):

1. Use a cluster bootstrap scheme to create B bootstrap samples, selecting J groups with replacement from the original sample.

2. Estimate $\hat{\Sigma}_B^*$, the between-level covariance matrix for each bootstrap sample.

3. Calculate $V = cov(\hat{\Sigma}_B^*)$ across the B bootstrap samples.

4. Estimate Γ_B, the asymptotic covariance matrix, as $\hat{\Gamma}_{BOOT} = JV$, where J

describes the number of groups in the sample.

The inverse of this estimated asymptotic covariance matrix, $\hat{\Gamma}^{-1}_{BOOT}$ can be used as a weight matrix to obtain parameter estimates in the ADF discrepancy function (Equation 3.7). Correspondingly, $\hat{\Gamma}^{-1}_{BOOT}$ can be used to obtain T_{ADF}, and T_{CADF} as given in Equations 3.36 and 3.38, and ADF standard errors (Equation 3.39). $\hat{\Gamma}_{BOOT}$ can also be used to obtain robust standard errors (Equation 3.42) the rescaled test statistic T_{RML} (Equation 3.40), and the residual based test statistics T_{RADF} and T_{CRADF} (Equations 3.43 and 3.44).

The non-parametric cluster bootstrap for three level models (students nested in teachers nested in schools) proceeds similarly:

1. Use a cluster bootstrap scheme to create B bootstrap samples, selecting K level-3 groups with replacement.

2. Estimate $\hat{\Sigma}^{*}_{WG}$, the between-subgroups covariance matrix for each bootstrap sample.

3. Calculate $V = cov(\hat{\Sigma}^{*}_{WG})$ across the B bootstrap samples.

4. Estimate Γ_{WG}, the asymptotic covariance matrix, as $\hat{\Gamma}_{BOOT} = JV$, where J describes the number of subgroups in the sample.

As in the case of the two-level cluster bootstrap, this estimated asymptotic covariance matrix can be used to obtain estimates of parameters and standard errors, as well as the ADF, rescaled and residual based test statistics T_{ADF}, T_{CADF}, T_{RML}, T_{RADF} and T_{CRADF}.

CHAPTER 5

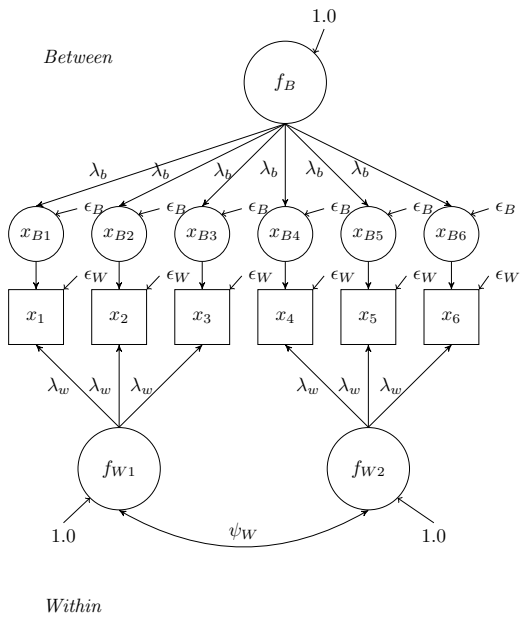
Methods

This dissertation investigates four issues that examine the performance of the segregating approach to multilevel factor analysis under real world conditions: 1) The efficiency of the segregating approach compared to the partially saturated model method in the estimation of parameters in two-level models. 2) The comparative performance of ADF estimation based on $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$ in the segregated analysis of $\hat{\Sigma}_B$ 3) The extension of the bootstrap-based method to three-level models. 4) The application of these bootstrap methods to a realistic dataset to investigate the dimensions of professional practice that are discernible in a state-wide student survey of instructional quality. Topics 1-3 are investigated using two simulation studies. Topic 4 is investigated through an application to an empirical data set. The simulation studies and empirical demonstration are described in detail in this chapter.

5.1 Simulation study 1: two-level factor analysis and the segregating method

The first simulation study examined the relative efficiency of the segregating approach, as compared to the partially saturated modeling method described in Section 3.5 (research topic 1), and examined the performance of ADF estimators based on $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$ under conditions likely to be encountered in realistic settings with student survey data (research topic 2). In order to do this, data was

Figure 5.1: Generating model for study 1



generated over a range of conditions of item intraclass correlations (ICCs), within group sample sizes, total numbers of groups, and model sizes. For all conditions, data were generated from multivariate normal distributions and a population model with two within-level factors and one between-level factor. This population model was selected because several sources suggest that the between-level factor structure is likely to be simpler than the structure at the within-level (e.g., Holfve-Sabel & Gustafsson, 2005; B. O. Muthén & Asparouhov, 2011). However, because the level-1 and level-2 covariance matrices were made independent by the design of the simulation, the generating model at level-1 should not influence the model at level-2 (e.g., Ryu & West, 2009). An illustrative path diagram with six items for this population model is given in Figure 5.1. The factor variances are set to 1, the between-level factor loadings, λ_B are all equal as are the uniquenesses ϵ_B. At the within level, all factor loadings (λ_W) are equal, as are the uniquenesses (ϵ_W). Note that ϵ_B may not equal ϵ_W, and λ_B may not equal λ_W. The covariance between the latent factors (ψ_W) was set to .3.

Each simulation condition consisted of 500 replications. For the bootstrap-based analyses, 500 bootstrap samples were used. Simulations were conducted using MPlus's Monte Carlo capabilities. For each of the replicated data sets, MPlus (L. K. Muthén & Muthén, 2010) and the MPlusAutomation package (Hallquist & Wiley, 2013) in R (R Core Team, 2013) were used to obtain saturated estimates of Σ_B and Σ_W. Once these two covariance matrices had been obtained, model parameters, standard errors and the test statistics T_{ADF}, T_{CADF}, T_{ML}, T_{RML}, T_{RADF} and T_{CRADF} were estimated in EQS (Bentler, 2006) using the REQS package (Mair & Wu, 2012) in R. In all cases, the correct model was fit to the simulated data.

5.1.1 Simulation conditions for study 1

Simulation conditions were selected in order to reflect the range of conditions that are commonly reported in survey-based research on classroom climate. Four simulation conditions were manipulated. These include: 1) the level-2 sample size, 2) the level-1 sample size (i.e. the level-1 units per level-2 unit), 3) the item ICCs, and 4) the size of the measurement model. A brief description of each of these conditions follows.

5.1.1.1 level-2 sample size

Three different level-2 sample sizes were included in this simulation: $J = 50$, $J = 100$, and $J = 200$. The range of level-2 sample sizes that were included in the simulation reflects the range of sample sizes reported in the applied literature. Sample sizes of around 50 classrooms are not uncommon (e.g., Holfve-Sabel & Gustafsson, 2005; Toland & De Ayala, 2005). Fauth et al. (2014) reported results based on 89 classrooms, and Kunter et al. (2008) reported results based on 323 classes.

5.1.1.2 level-1 sample size

Three different level-1 sample sizes were included in this simulation: $n = 10$, $n = 30$, and $n = 50$. Because class sizes tend to be directly influenced by national, state or local policies, there is less reported variation in the number of students within each classroom in the applied literature. These sizes range from around 12 student per class (Kunter et al., 2008) to approximately 25 students per class (Holfve-Sabel & Gustafsson, 2005).

5.1.1.3 Item intraclass correlations

Four different ICC conditions were included in simulation: $ICC = .05$, $ICC = .10$, $ICC = .26$, and $ICC = .50$. Marsh et al. (2012) noted that, generally speaking, item ICCs for climate variables are "often less than .1 and rarely greater than .3" (p. 115). This is consistent with reported item ICCs in the applied literature. Toland and De Ayala (2005), for example, reported ICCs that range from .06 to .43. den Brok, Stahl, and Brekelmans (2004) reported ICCs ranging from .06 to .24.

ICCs were varied across simulation conditions by altering the within-level factor loadings and uniquenesses. For all conditions, $\lambda_B = .7$and $\epsilon_B = .51$. For the $ICC = .50$ condition, $\lambda_W = .7$ and $\epsilon_W = .51$. For the $ICC = .26$ condition, $\lambda_W = 1.41$ and $\epsilon_W = 2.00$. For the $ICC = .10$ condition, $\lambda_W = 4.59$ and $\epsilon_W = 2.10$. For the $ICC = .05$ condition, $\lambda_W = 3.08$ and $\epsilon_W = 9.50$. This was done in order to keep the proportion of reliable variance at the between-level (i.e., the variance in the latent means that is accounted for by the factor) constant across simulation conditions. Previous simulation work (Hox et al., 2010) has indicated that changing indicator reliability across simulation conditions can make the interpretation of results difficult.

5.1.1.4 Model size

Two different model sizes were used. One with 6 observed variables and 9 degrees of freedom, and one with 12 observed variables and 54 degrees of freedom. Because classroom climate is a complex and multidimensional construct, it is typically assessed using a large number of survey items. This leads to measurement models that are relatively large. The smaller measurement models reported in the literature contain around 25 degrees of freedom (Kunter et al., 2008). den Brok et al. (2004) tested a variety of measurement models with approximately 60 degrees of freedom. Toland and De Ayala (2005), Fauth et al. (2014) and Holfve-Sabel and Gustafsson (2005) used between-level factor models with over 150 degrees of freedom.
In total, the simulation contained in $3 \times 3 \times 4 \times 2 = 72$ conditions and a total of $72,000$ replications ($36,000$ using GEE methods, and $36,000$ using bootstrap-based methods). While certain constellations of conditions may be unlikely to occur in practice (i.e. many, large classrooms, high ICCs and a small model), the inclusion of conditions across this range allows for a more comprehensive investigation.

5.1.2 Measures of performance for simulation study 1

Because the data were generated under the assumption of multivariate normality with known population parameters, it is possible to describe Γ_B exactly. Specifically, results in Yuan and Bentler (2006) imply that Γ_B is given by the inverse of the Fisher Information:

$$\Gamma_B = W_J^{-1} + \frac{1}{n^2}\left((n-1)W_W\right)^{-1} \tag{5.1}$$

Note that:

$$W_W = .5D_p^T[\Sigma_W^{-1} \otimes \Sigma_W^{-1}]D_p \tag{5.2}$$

where Σ_W is the within-groups covariance matrix as defined in Equation (3.22). Also note that:

$$W_J = .5D_p^T[\Sigma_J^{-1} \otimes \Sigma_J^{-1}]D_p \tag{5.3}$$

where $\Sigma_J = \Sigma_B + \frac{1}{n}\Sigma_W$. In this paper, Γ_B given in Equation 5.1 is referred to as Γ_{FISHER}.

Γ_{FISHER}^{-1} can be used in the generalized least squares discrepancy function in Equation 3.4 to obtain consistent estimates of parameters and a chi-square test statistic. This test statistic, called T_{FISHER} in this dissertation, is asymptotically distributed as a central chi-square variate under the null hypothesis. In addition, Γ_{FISHER} can be used to calculate the correct standard errors (based on Equation 3.39). Because of this, Γ_{FISHER} provides a good basis of comparison for the performance of the standard errors and for the asymptotic performance of the test statistics estimated using $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$ (Yung & Bentler, 1994). In addition, the accuracy of the estimation of $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$ themselves can be appraised by comparison with Γ_{FISHER} . Research topics 1 and 2 will be investigated through the following sources of information:

1. Relative efficiency. In order to compare the efficiency of the partially saturated model method and segregating approach, the ratio of the mean square errors of the parameter estimates was compared (e.g., Hoel, Port, & Stone, 1971):

$$e(\hat{\theta}_{SAT}, \hat{\theta}_{SEG}) = \frac{E(\hat{\theta}_{SEG} - \theta)^2}{E(\hat{\theta}_{SAT} - \theta)^2} \tag{5.4}$$

   If $e < 1$, the segregating approach would be preferable to the partially saturated model method. e can be used to compare the relative efficiency of parameter estimation using ADF and ML estimators in conjunction with the segregating approach to the estimation of parameters using the ML estimator in conjunction with the partially saturated model method.

2. Parameter bias. Parameter bias is given by $E(\hat{\theta} - \theta)$ (e.g., Flury, 1997). Bias was monitored for all estimated factor loadings in all conditions for all estimators.

3. Mean square error. Mean square error (MSE) is given by: $E(\hat{\theta} - \theta)^2$. MSE provides a way to assess the accuracy with which parameters are estimated. Mean square error is the sum of the parameter variance and squared bias (e.g., Flury, 1997). So for unbiased parameter estimates, MSE quantifies the sampling variance of those parameters (e.g., Lüdtke, Marsh, Robitzsch, & Trautwein, 2011). An estimator is considered consistent if $\lim_{n\to\infty}(MSE(\theta)) = 0$ (e.g., Flury, 1997). MSE was monitored for factor loadings for each estimator across all conditions.

4. Standard errors. The accuracy and consistency of the standard errors was assessed by monitoring the estimated standard errors for each simulation condition. The performance of standard errors obtained either through ML estimation, or through ADF methods (based either on the $\hat{\Gamma}_{GEE}$ or on $\hat{\Gamma}_{BOOT}$) can be compared to standard errors based on Γ_{FISHER}, which are known to be correct under normality. In this analysis, standard error estimates were compared based on compared based on the square of the L2 norm (Horn & Johnson, 2012):

$$D^2 = ||\hat{SE} - SE_{FISHER}||^2 \tag{5.5}$$

   The smaller the value of D^2, the closer the estimated standard error is to the correct standard error given by SE_{FISHER}. Analogous to the case of MSE, if $D^2 \to 0$ as sample size increases, the ML or ADF estimator can be called consistent (e.g. Yuan & Hayashi, 2006).

5. Type I error rates. For each estimated test statistic, an empirical model rejection rate was calculated. For the purpose of this study, the rejection rate was calculated at the nominal $\alpha = .05$ level. Because it is expected

that the empirical error rates will differ somewhat from the nominal rate, an acceptable empirical error rate is taken as one that falls in the interval [.028, .079], the estimated 2-sided 99% adjusted Wald confidence interval (e.g., Agresti & Coull, 1998).

6. Test statistic means and standard deviations. Means and standard deviations of the test statistics were estimated for each condition, based on fitting the correct model to each replicated data set. For a well-behaved test statistic (e.g. Curran et al., 1996), the observed means and standard deviations should be close to the theoretical values. For a central chi square distribution, the mean is given by d, the degrees of freedom, and the variance is given by $2d$.

7. Q-Q Plots. Presented for each estimated test statistic (T_{ADF}, T_{CADF}, T_{ML}, T_{RML} , T_{RADF} , T_{CRADF} , and T_{FISHER}). Q-Q plots of test statistic distributions help visualize the empirical sampling distribution of the test statistics, and provide information about the overall distribution of a test statistic. Q-Q plots are particularly helpful at showing deviations from expected statistic distribution in the tails (Gnanadesikan, 1977).

8. Asymptotic covariance matrices. In addition to test statistics, $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$ were monitored, and compared through their square distances from Γ_{FISHER}: $D^2 = ||vech(\hat{\Gamma}) - vech(\Gamma_{FISHER})||^2$. The smaller the value of D^2, the closer the estimated asymptotic covariance matrix is to the correct matrix given by Γ_{FISHER}.

5.2 Simulation study 2: three-level models

Study 2 addresses the third research question, regarding the expansion of the segregating approach to three-level models. There is very little literature that explicitly considers models with three levels. While several sources (Goldstein,

2003; Longford & Muthén, 1992; Yau et al., 1993; Yuan & Bentler, 2007) describe theoretically the extension of two-level factor analytic frameworks to three levels and beyond, only Yau et al. (1993) provided an empirical demonstration of this method. As far as this author knows, no published studies have applied three-level factor analysis techniques to the analysis of student surveys of professional practice. Ren et al. (2010) demonstrated the three-level cluster bootstrap, but did not apply the technique on multivariate data.

This dissertation uses a simulation study to provide an empirical verification that the cluster bootstrap can be used to extend estimation methods to three-level data sets (research topic 3). Because the possible combinations of simulation conditions for three-level models are exponentially larger than those with two models, simulation study 2 is significantly smaller than simulation study 1, and focuses specifically on one set of conditions likely to be encountered in student survey research or in the use of student ratings to evaluate professional practice. Simulation study 2 explored the estimation of parameters, and the behavior of test statistics in the segregated analysis of $\hat{\Sigma}_{WG}$, the between subgroups level covariance matrix. In simulation study 2, ML estimation was used, as this estimator showed the best performance in simulation study 1, and only the test statistics computed under ML estimation— T_{RADF}, T_{CRADF} and T_{RML}—were explored in simulation study 2.

Data were generated from multivariate normal distributions and a population model with one level-1 factor, one level-2 factor, and one level-3 factor. While the generating model in simulation study 1 contained three factors at level-1 (the within level, see Figure 5.1), the model in simulation study 2 was made simpler in order to decrease the computational demands of the simulation. It was also hypothesized that, because the level-1, level-2 and level-3 covariance matrices were made independent by the design of the simulation, the generating models at level-1 and level-3 should not influence the model at level-2 (e.g., Ryu & West, 2009). An

illustrative three-level path diagram with six items for this population model is given in Figure 5.2. The factor variances are set to 1, and within each level, the factor loadings are all equal, as are the uniquenesses. However, $\lambda_B \neq \lambda_{WG} \neq \lambda_{WGK}$, and $\epsilon_B \neq \epsilon_{WG} \neq \epsilon_{WGK}$. Each simulation condition consisted of 250 replications. Fewer replications were included in this study because the cluster bootstrap and the estimation of the covariance matrices was more computationally intensive than in study 1. Each replication took between 5 and 7 minutes on a quad-core i7 processor (a little over 1 day in total) . Simulations were conducted using MPluss Monte Carlo capabilities. For each of the replicated data sets, MPlus (L. K. Muthén & Muthén, 2010) and the MPlusAutomation package (Hallquist & Wiley, 2013) in R (R Core Team, 2013) were used to obtain an estimate of Σ_{WG}. Model parameters and test statistics were estimated in EQS (Bentler, 2006) using the REQS package (Mair & Wu, 2012) in R. The correct model was fit to each simulated data set.

5.2.1 Simulation conditions for study 2

Because only a single set of conditions was included in simulation study 2, the model parameters and sample sizes were selected based on the data set that will be used in the empirical illustration. Specifically, the first six items from the Opportunity to Learn (OTL) Survey, a student survey of instructional practice that has recently been implemented in the state of New Mexico, were used as observed indicators. Level-1, level-2 and level-3 sample sizes were also inspired by the classroom and school configurations in the OTL data set. Complete details on the OTL survey are provided in the next section. For the purposes of simulation study 2, the OTL survey was used only to provide a set of conditions that would approximate real-world conditions. Using the OTL survey in this way helps to improve the generalizability of the findings from the single simulation condition to

subsequent real-world applications. The population factor loadings were obtained in the following way: 1) The three-level model given in Figure 5.2 was fit to six variables from the OTL survey data. Only six variables were used because this resulted in a between subgroup (level-2) measurement model with 9 degrees of freedom, the same as the small model conditions used in simulation study 1. 2) Model parameters were estimated, and then used to inform the population parameters in a data generating model, which was also based on the model given in Figure 5.2.

Specifically, the factor loadings and uniquenesses obtained by fitting the model in Figure 5.2 were $\lambda_{WGK} = .528$, $\lambda_{WG} = .271$ $\lambda_B = .177$. $\epsilon_{WGK} = 1.25$, $\epsilon_{WG} = .05$, and $\epsilon_B = .01$. This pattern of factor loadings and unique variances implies level-2 ICCs of .07, and level-3 ICCs of approximately .025. A total of 110 groups were included in the sample, with 17,600 individuals. This translated to a level-1 sample size of 20 (20 individuals per subgroup), and a level-2 sample size of 8 (8 subgroups per group). Performance of the cluster bootstrap approach in the three-level case will be monitored using measures of performance analogous to those used in Simulation 1:

1. Parameter bias. Bias was monitored for all estimated factor loadings in all conditions.

2. Mean square error. MSE was monitored for all estimated factor loadings in all conditions.

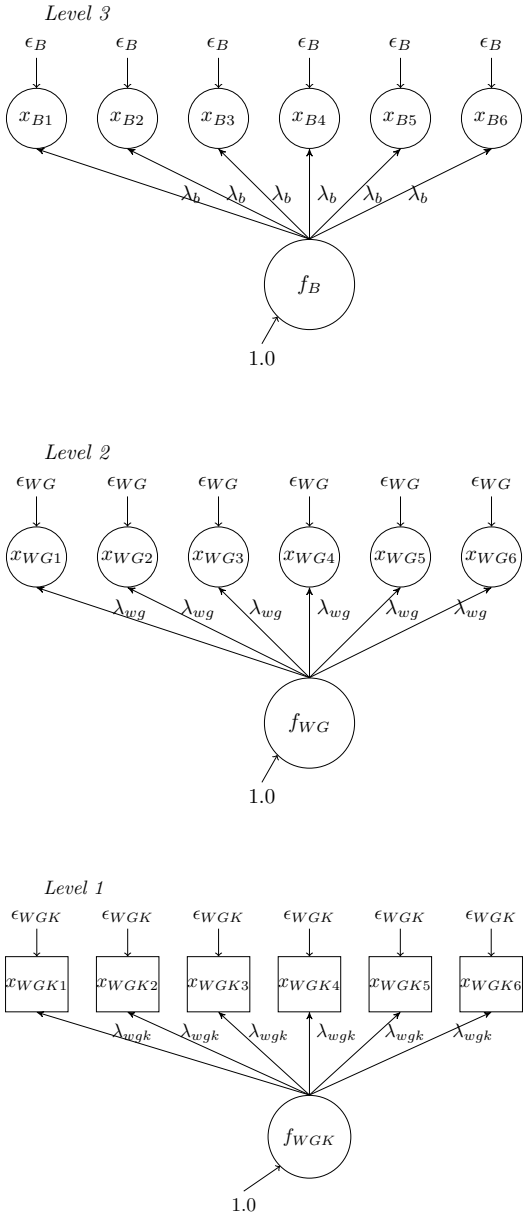
3. Type I error rates. For each estimated test statistic, an empirical model rejection rate was calculated. For the purpose of this study, the rejection rate was calculated at the nominal $\alpha = .05$ level. Because it is expected that the empirical error rates will differ somewhat from the nominal rate, an acceptable empirical error rate is taken as one that falls in the interval [.020, .095], the estimated 2-sided 99% adjusted Wald confidence interval

(e.g., Agresti & Coull, 1998).

4. Test statistic means and standard deviations. Means and standard deviations of the test statistics were estimated for each condition, based on fitting the correct model to each replicated data set. For a well-behaved test statistic (e.g. Curran et al., 1996), the observed means and standard deviations should be close to the theoretical values. For a central chi square distribution, the mean is given by d, the degrees of freedom, and the variance is given by $2d$.

5. Q-Q Plots. Presented for each estimated test statistic (T_{RML}, T_{RADF}, T_{CRADF}). Q-Q plots of test statistic distributions help visualize the empirical sampling distribution of the test statistics, and provide information about the overall distribution of a test statistic. Q-Q plots are particularly helpful at showing deviations from expected statistic distribution in the tails (Gnanadesikan, 1977).

Figure 5.2: Generating model for study 2



5.3 Empirical illustration: the New Mexico Opportunity to Learn survey

In order to investigate the fourth research topic, statewide data from the New Mexico Opportunity to Learn student survey was used to illustrate 1) how the segregated approach can be used to investigate the dimensions of instructional practice that are discernible based on aggregated student responses and 2) how those factors relate to outcomes of interest. Specifically, this empirical illustration builds on the two simulation studies by demonstrating how the cluster bootstrap can be applied to a real data set, where the population model is unknown.

5.3.1 Teacher evaluation in New Mexico and NMTEACH

In 2011, the New Mexico Effective Teaching Task Force issued a set of recommendations about policies aimed at "recruiting, retaining and rewarding New Mexicos most effective teachers and school leaders" (New Mexico Effective Teaching Task Force, 2011, p. 4). The recommendations include a new framework for teacher evaluation (NMTEACH) based on three sources of evidence: value added (VAM) estimates of student achievement growth (50% of the overall evaluation); ratings of practice based on classroom observations (25%); and locally adopted measures (25%). The latter measures are left to the discretion of local education agencies, but the task force does describe specifically that these multiple measure may include "portfolios of teacher and student work, surveys of parents or students, or other research-based measures proven to demonstrate or correlate to student learning gains" (New Mexico Effective Teaching Task Force, 2011, p. 5). A prototype of a student survey to be offered as part of NMTEACH, was first administered in the 2011-12 academic year with the state's Standards Based Assessment.

The Opportunity to Learn (OTL) Survey is a 10 item survey designed to measure the quality of instruction and the school environment. Different versions of the sur-

vey are administered in elementary (grades 3-5) and middle and high school (grades 6-12). Each item is scored on a 6-point scale, from 0 to 5, where the categories are $0 = never$, $1 = hardly\ ever$, $2 = sometimes$, $3 = usually$, $4 = almost\ always$, and $5 = always$.

Data used in this study were collected in the 2012-2013 administration of the OTL survey. In the 2012-2013 school year, there were 338,223 K-12 students enrolled in the New Mexico public school system. This analysis focuses only on the early grades version of the survey, where student raters are uniquely nested within a single teacher. The dataset was restricted to include students that could be uniquely linked with teachers, and to teachers with class sizes of 10 or more. These students were enrolled in 997 school sites. This included 76,865 students enrolled in grades 3-5. The dataset used in this study contained 63,064 grade 3-5 students with complete OTL Survey responses. This represents approximately 82% of the total student enrollment statewide in grades 3-5. These students were nested within 3278 classrooms in 443 schools. Table 5.1 displays demographic information for these survey respondents.

Table 5.1: New Mexico student demographics (grades 3-5)

	N	Percent
White	15,065	24.7%
Black or African American	1,377	2.2%
Hispanic	38,812	61.5%
Asian	946	1.5%
Native American	6,324	10.0%
Female	31,581	50.1%
Free or Reduced Lunch	46,794	74.2%
Students with Disabilities	7,083	11.2%
English Language Learner	12,269	19.5%[a]
Total	63,064	

Note: [a] based on 63,013 responses

5.3.2 Analysis Approach

The empirical illustration contains two analyses, demonstrating how the segregating method may be used to investigate factorial structure, and to investigate how factors relate to outcomes of interest. First, the segregating method was used to determine the dimensions of instructional practice that are discernable based on aggregated students responses in the OTL survey. Second, the aggregated survey variables were used to predict student achievement growth in math and reading. The specifics of these analyses are described in the sections that follow.

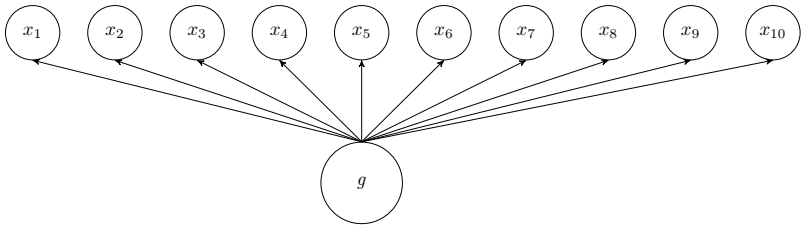
5.3.2.1 Illustrating how the segregating method can be used to determine between-teacher covariance structure

Because the elementary student data was hierarchically structured with three levels—student, classrooms, and schools—the segregating approach was used to analyze the between-classroom covariance matrix. Maximum likelihood estimation was used in conjunction with the cluster bootstrap estimate of the asymptotic covariance matrix, so that rescaled and residual-based test statistics could be used in model appraisal, and robust standard errors could be used to make inferences about model parameters.

Opportunity to Learn is a broadly defined construct that has often been found to be a strong predictor of student achievement (e.g., Brophy & Good, 1986; Guiton & Oakes, 1995; Murphy, 1988; Saxe, Gearhart, & Seltzer, 1999; Walker & Schaffarzick, 1974; Wang, 1998; K. B. Wu, Goldschmidt, Boscardin, & Sankar, 2009). The OTL survey focuses on one specific aspect of Opportunity to Learn, the "quality of instructional delivery", defined as the variety of teaching strategies teachers use in order to meet the educational needs of all students. (Brophy & Good, 1986; Stevens & Grymes, 1993; Stigler & Stevenson, 1992).

While "Quality of Instructional Delivery" could be thought of as a unidimensional

Figure 5.3: Unidimensional model for Opportunity to Learn survey

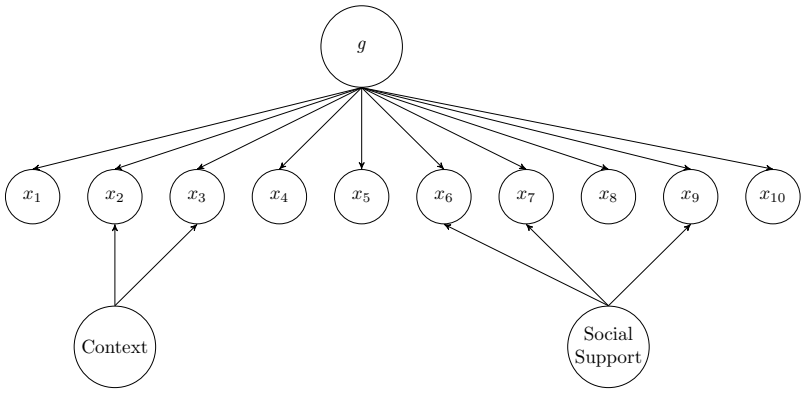


construct, it is also possible that there are discernible subdimensions that exist in addition to overall quality of instruction. For example, Kunter et al. (2008) identified the concept of "social support", which refers to the extent to which a teacher creates "a supportive social environment in which students receive personal guidance and feel personally valued" (p. 471). Another dimension that is often discussed in relation to instructional practice concerns how teachers communicate the "value interest and intrinsic reasons inherent in schoolwork" (Patrick, Turner, Meyer, & Midgley, 2003, p.1525). In other words, the extent to which teachers provide students with a sense of trajectory—where the content comes from, where it goes next, and how it connects to other bodies of knowledge.

Because there are two different theoretically possible positions here—namely that instructional quality is unidimensional, or that there exist discernible subdimensions of instructional quality—two different *a priori* models were fit to the OTL survey data (MacCallum, Roznowski, Mar, & Reith, 1994). These include a unidimensional model and a bifactor model. These models are described below. For each model, only the between teacher level covariance structure is described. For simplicity, error variance and factor variances have been omitted from the diagrams. The unidimensional model (Figure 5.3) would support a theory that students, on average are able to discern differences in overall instructional practice across teachers, but are unable to make finer distinctions about specific aspects of teacher practice. The bifactor model (Holzinger & Swineford, 1937) (Figure 5.4) can be used to examine the possibility that there is a common factor, and also additional

specific factors "caused by parcels of items tapping similar aspects of the trait" (Reise, Moore, & Haviland, 2010, p. 549). The model used in this analysis is an

Figure 5.4: Bifactor model for Opportunity to Learn survey



"incomplete" bifactor model (Chen, West, & Sousa, 2006) and it differs from a canonical bifactor model in that not all of the items load onto the general factor and one other specific factor. In this case of the OTL survey, this model suggests that there is one underlying general trait (instructional practice), but that there are subgroups of items "tapping on similar aspects of this trait" (p. 549)—specifically, the three items that are about social support and the extent to which "students receive personal guidance" (Kunter et al., 2008, p. 471) (*social support*), and two items that are about the extent to which teachers provide students with a sense of trajectory (*context*). It should also be noted that generally speaking, when the specific factors have only two indicators, the model will not be identified without additional constraints (see Devena, Gay, and Watkins (2013) for an illustrative example).

Model selection was based on chi-square tests and three fit indices: the Root Mean Square Error of Approximation (RMSEA) (Steiger & Lind, 1980), the Comparative Fit Index (CFI) (Bentler, 1990) and the average absolute standardized residual. Research has suggested these indices can be helpful for assessing model fit (Browne, MacCallum, Kim, Andersen, & Glaser, 2002; Tomarken & Waller, 2003).

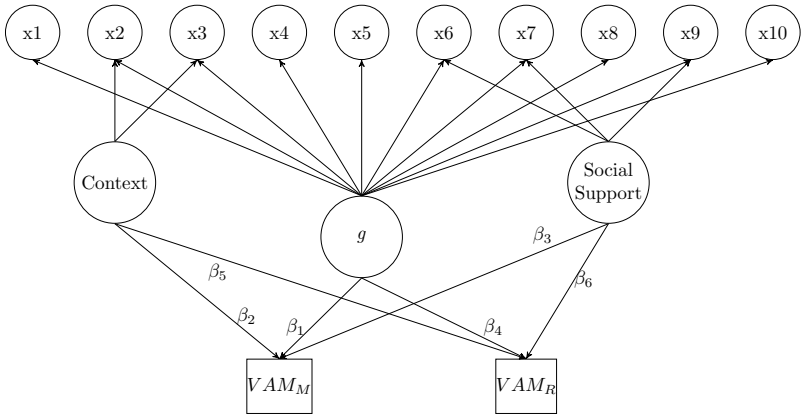
5.3.2.2 Illustrating how the segregating method can be used to investigate relationships with external variables

In order to illustrate how the segregating method may be used to investigate relationships between teacher-level variables and external criteria, VAM scores were computed for each teacher in both math and reading, based on student performance on the New Mexico Standards Based Assessments (SBAs). This analysis differs from the previous analyses, in that the VAM scores do not have any within-class variation. In this sense, this analysis is of the type described in Section 3.7, and given in Equation 3.59, where VAM scores are measured at the teacher level, and contain only between classroom and between school variance components. As was described in Section 3.7, even when variables with level restricted variation are included, estimators such as those given in Equation 3.54 can still be used to obtain consistent estimates of the between-classroom level covariance matrix.

VAM scores were computed following the methodology outlined in T. J. Kane et al. (2013). The estimated VAM scores describe the extent to which students of a particular teacher, on average, performed relative to similarly situated students (T. J. Kane et al., 2013).

The estimated VAM scores were used as external variables in a latent variable model with the general and domain specific factors acting as predictors (Figure 5.5). In this model, β_1 and β_4 refer to the relationship between general instructional practice and VAM scores. β_2, β_3, β_5 and β_6 describe the extent to which the specific factors *context* and *social support* predict VAM scores above and beyond the general instructional practice factor.

Figure 5.5: Bifactor model for predicting estimated teacher value added scores



CHAPTER 6

Simulation Study Results

This chapter presents and discusses results from the simulation studies investigating the first three research topics. First, it presents results regarding the relative efficiency of the segregating approach compared to the partially saturated model method in the estimation of parameters in two-level models. Second, it presents results on the comparative performance of GEE-based ADF, cluster bootstrap-based ADF and ML estimators in the segregated analysis of $\hat{\Sigma}_B$. Third, the cluster bootstrap is extended to three level models, and applied to the segregated analysis of the between-subgroups covariance matrix $\hat{\Sigma}_{WG}$. The results are organized by research question and the chapter closes with a synthesis and a discussion of the results.

6.1 The relative efficiency of the segregating approach

Tables 6.1 and 6.2 display the relative efficiency of the parameter estimates obtained from the segregating approach and the partially saturated model method across all simulation conditions. The relative efficiency was determined by the ratio of the mean square errors of the parameter estimates from the partially saturated model method and the segregating approach. If this ratio is less than 1, the segregating approach would be preferable to the partially saturated model method. If the ratio is greater than 1, the partially saturated model method would be preferable to the segregating approach.

Because the segregating approach used three different estimators (ML estimation,

Table 6.1: Efficiency of the segregating method, relative to the partially saturated model method, $df = 9$

		Group Sizes								
		10			30			50		
		ML	GEE	Boot	ML	GEE	Boot	ML	GEE	Boot
$J = 200$	$ICC = .50$	1.00	*1.09*	*1.07*	1.00	*1.09*	*1.07*	1.00	*1.12*	*1.11*
	$ICC = .26$	1.00	*1.10*	*1.08*	1.00	*1.09*	*1.07*	1.00	*1.13*	*1.10*
	$ICC = .10$	1.02	*1.22*	*1.13*	1.01	*1.11*	*1.10*	1.00	*1.13*	*1.10*
	$ICC = .05$	**0.90**	*1.29*	*1.10*	1.02	*1.22*	*1.15*	1.01	*1.14*	*1.10*
$J = 100$	$ICC = .50$	1.00	*1.26*	*1.24*	1.00	*1.16*	*1.16*	1.00	*1.20*	*1.19*
	$ICC = .26$	1.00	*1.29*	*1.28*	1.00	*1.19*	*1.17*	1.00	*1.22*	*1.20*
	$ICC = .10$	**0.83**	*1.28*	*1.12*	1.01	*1.25*	*1.24*	1.00	*1.26*	*1.24*
	$ICC = .05$	**0.48**	**0.80**	**0.69**	1.04	*1.52*	*1.33*	*1.01*	*1.40*	*1.30*
$J = 50$	$ICC = .50$	1.00	*1.82*	*1.72*	1.00	*1.52*	*1.57*	*1.00*	*1.53*	*1.54*
	$ICC = .26$	1.00	*1.67*	*1.77*	1.00	*1.60*	*1.67*	1.00	*1.58*	*1.54*
	$ICC = .10$	**0.14**	**0.25**	**0.24**	1.02	*1.85*	*1.71*	1.00	*1.62*	*1.57*
	$ICC = .05$	**0.13**	**0.21**	**0.17**	**0.42**	**0.80**	**0.70**	**0.88**	*1.44*	*1.41*

and ADF estimation with two different estimates of the asymptotic covariance matrix, $\hat{\Gamma}_{GEE}$, and $\hat{\Gamma}_{BOOT}$), three different ratios are reported for each simulation condition. In both tables, conditions where the partially saturated model method was considerably more efficient (at least 5% more efficient) are shown in italics, and conditions where the segregating approach is more efficient are shown in bold. Table 6.1 displays results for the small model size condition $df = 9$. These results suggest that there is, in general, no loss of efficiency that comes from using the segregating method in conjunction with maximum likelihood estimation, because most of the ratios are less than or equal to 1 across all model conditions. Supporting the hypotheses of Yuan and Bentler (2007), there is even a gain in efficiency for the segregating method as the ICCs get smaller and the group sizes get smaller. At $ICC = .10$ and $ICC = .05$ with a small number of groups and only 10 individuals in each group, the segregating method is far more efficient than the partially saturated model method, with ratios as small as .13 ($ICC = .05$, $n = 10$, $J = 50$).

The ADF estimators are, in general, less efficient than the partially saturated model method, and the cluster bootstrap produces slightly less efficient estimates

Table 6.2: Efficiency of the segregating method, relative to the partially saturated model method, $df = 54$

		Group Sizes								
		10			30			50		
		ML	GEE	Boot	ML	GEE	Boot	ML	GEE	Boot
$J = 200$	$ICC = .50$	1.00	*1.82*	*1.63*	1.00	*1.85*	*1.65*	1.00	*1.80*	*1.60*
	$ICC = .26$	1.00	*2.20*	*1.85*	1.00	*1.88*	*1.69*	1.00	*1.82*	*1.61*
	$ICC = .10$	0.96	*3.32*	*2.39*	1.00	*2.05*	*1.76*	1.00	*1.86*	*1.72*
	$ICC = .05$	**0.70**	*2.12*	*2.88*	0.99	*2.51*	*1.96*	0.99	*2.10*	*1.84*
$J = 100$	$ICC = .50$	1.00	*4.54*	*4.17*	1.00	*4.34*	*4.29*	1.00	*4.10*	*3.83*
	$ICC = .26$	0.99	*4.75*	*4.52*	1.00	*4.58*	*4.46*	1.00	*3.89*	*4.02*
	$ICC = .10$	**0.72**	*3.27*	*2.94*	0.99	*5.05*	*5.07*	1.00	*4.14*	*4.07*
	$ICC = .05$	**0.51**	*1.72*	*1.56*	0.9	*4.58*	*4.42*	0.99	*4.57*	*4.28*
$J = 50$	$ICC = .50$	1.00			1.00			1.00		
	$ICC = .26$	0.96			1.00			1.00		
	$ICC = .10$	**0.64**			0.98			0.99		
	$ICC = .05$	**0.55**			**0.70**			**0.86**		

than the GEE-based estimator. This is not unsurprising, as, under normality, ML estimation is expected to be asymptotically most efficient. However, what is surprising is that at low ICCs ($ICC = .05$), small group sizes, and a small number of groups, both segregated ADF approaches are more efficient than using ML estimation in the partially saturated model method. This pattern of results can also be seen in the larger models ($df = 54$), which are displayed in Table 6.2. Again, there is no loss of efficiency that comes from using the segregating method in conjunction with maximum likelihood estimation. For all ICC conditions, group sizes, and numbers of groups, the segregating approach is at least as efficient as the partially saturated model method. And, while there is a gain in efficiency for the segregating approach at small ICCs and group sizes, the relative performance of the two methods is much more comparable.

The ADF estimators, are, in general, far less efficient than the saturating approach, and the ADF approach employing $\hat{\Gamma}_{GEE}$ appears slightly more efficient than the ADF approach using $\hat{\Gamma}_{BOOT}$. Unlike the case with smaller models, ML estimation in the partially saturated model method is more efficient than the segregating method at all ICC and within-group sample size conditions. Note that ADF-

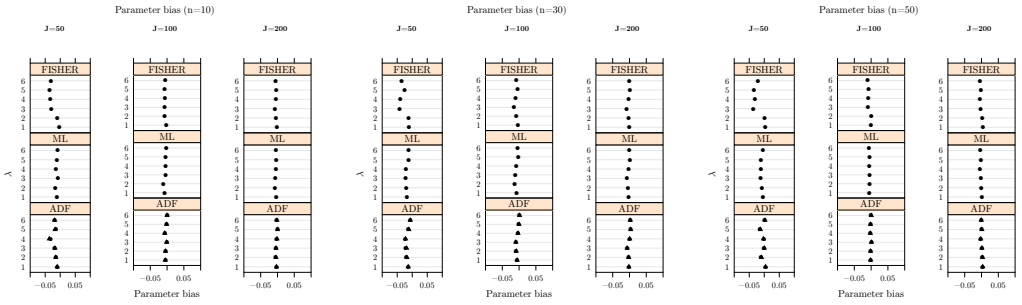
estimation was not possible with 50 groups and the larger models ($df = 54$), because $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{GEE}$ are, in general, not invertible under these conditions, and estimates are not computed by EQS.

6.2 Comparative performance of ADF and ML estimators in the segregated analysis of $\hat{\Sigma}_B$

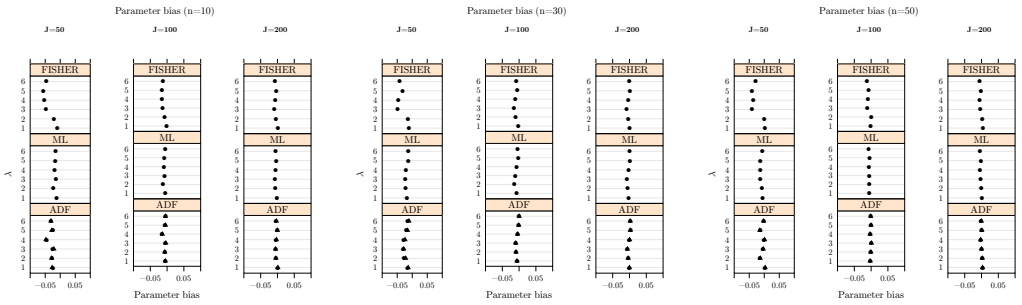
6.2.1 Parameter bias

Figures 6.1 and 6.2 present plots of the estimated parameter bias based on four different estimators, so that the performance of the GEE and cluster bootstrap based approaches can be assessed and compared. The upper panel of each plot displays parameter bias based on estimation using Γ_{FISHER}. Recall that Γ_{FISHER} is correctly specified under normality, and so parameter estimates based on Γ_{FISHER} should be consistent. The second panel displays parameter bias for the ML estimator. Note that the use of the bootstrap does not influence the ML parameter estimates. This is because the ML discrepancy function (Equation 3.12) uses only a model implied matrix and the sample covariance matrix to estimate optimal model parameters. The third panel displays parameter bias based on ADF estimation using $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$, the GEE and cluster bootstrap estimates of Γ_B. The GEE-based estimates are represented by black circles (•), and the cluster bootstrap based estimates are represented by black triangles (▲). Figure 6.1 displays only information for the small models ($df = 9$), and Figure 6.2 displays information for the large models ($df = 54$). If parameters were unbiased, the dots in each plot panel should be in a straight vertical line above the value of 0 on the horizontal axis. The patterns of bias are similar for the ADF and ML estimators. There is a slight negative bias in parameter estimates for all estimators when either: 1) level-2 sample sizes are small 2) within group sample sizes are small 3) ICCs are low. However, as the number of groups increases, this negative bias disappears for

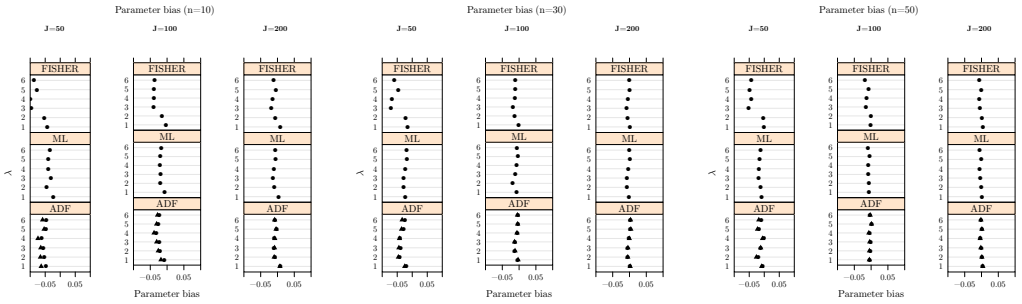
Figure 6.1: Parameter bias by estimator: $df = 9$



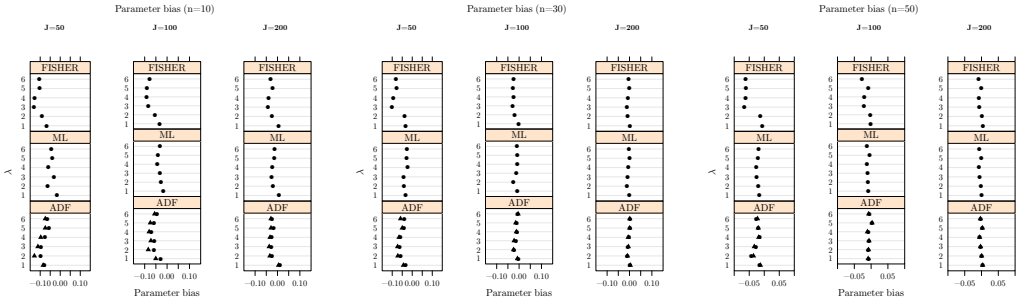
(a) $ICC = .50$ $n = 10$ (b) $ICC = .50$ $n = 30$ (c) $ICC = .50$ $n = 50$



(d) $ICC = .26$ $n = 10$ (e) $ICC = .26$ $n = 30$ (f) $ICC = .26$ $n = 50$



(g) $ICC = .10$ $n = 10$ (h) $ICC = .10$ $n = 30$ (i) $ICC = .10$ $n = 50$



(j) $ICC = .05$ $n = 10$ (k) $ICC = .05$ $n = 30$ (l) $ICC = .05$ $n = 50$

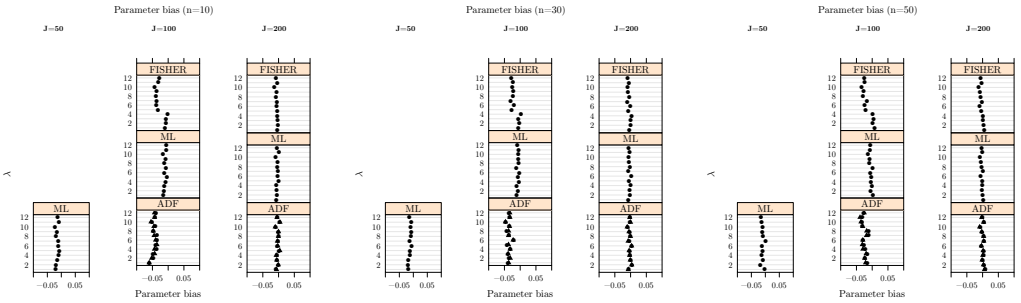
Legend: ▲ =Bootstrap, • =GEE

all level-1 sample size and ICC conditions. The trend across the plots in Figure 6.1 demonstrates that for sufficient numbers of groups, all of the estimators yield unbiased parameter estimates. Additionally, there is little difference between using ADF estimation in conjunction with $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$, showing that the cluster bootstrap based estimator can yield unbiased parameter estimates for sufficiently large level-2 sample sizes. Figure 6.2 shows parameter bias plots for the larger models ($df = 54$). Note that ADF estimation was not possible with 50 groups and the larger models ($df = 54$), because $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$ are, in general, not invertible under these condition and ADF estimation is not available in EQS.
For larger models, patterns of parameter bias are similar to the small models ($df = 9$). Specifically, as level-2 sample size increases, parameter bias decreases and becomes negligible for all estimation methods. However, there are slightly larger differences between the bootstrap and GEE-based ADF estimators at smaller sample sizes ($J = 100$), implying that the GEE-based estimator may outperform the bootstrap-based estimator for large models and small level 2 sample sizes. These differences are also more extreme for smaller ICCs. That is, for $ICC = .50$, the difference between ADF and bootstrap-based estimators is relatively small (Figure 6.2a). But for $ICC = .10$, the difference is larger (Figure 6.2g), and for $ICC = .05$, the difference is quite pronounced (Figure 6.2).

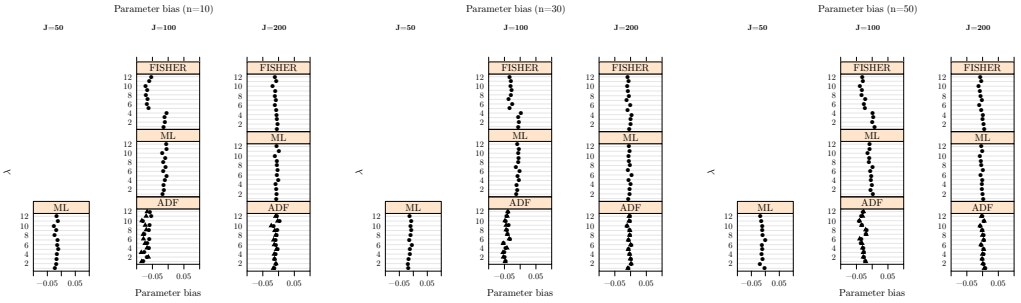
6.2.2 Variability of parameter estimates

Figures 6.3 and 6.4 present plots of the overall accuracy and sampling variability of the parameter estimates for four different estimators. Accuracy and variability of the parameter estimates was quantified using the mean square error (MSE). Mean square error can be thought of as the sum of variance and square bias. Thus, for unbiased parameters, MSE quantifies the sampling variability of the parameter estimates. Estimators are called consistent if $\lim_{n\to\infty}(MSE(\theta)) = 0$. The upper panel of each plot displays MSE based on estimation using Γ_{FISHER}. The second

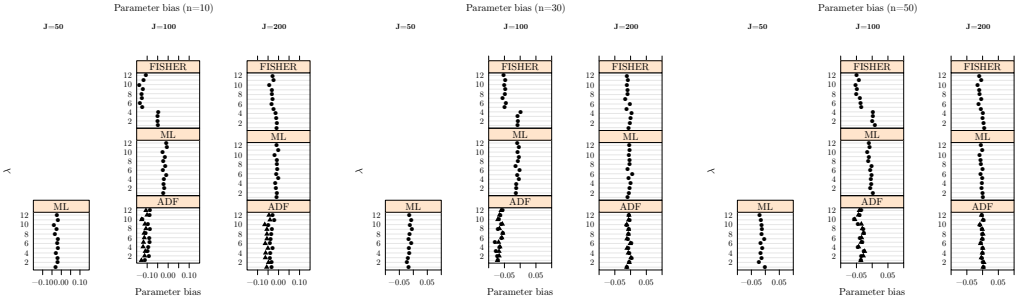
Figure 6.2: Parameter bias by estimator: $df = 54$



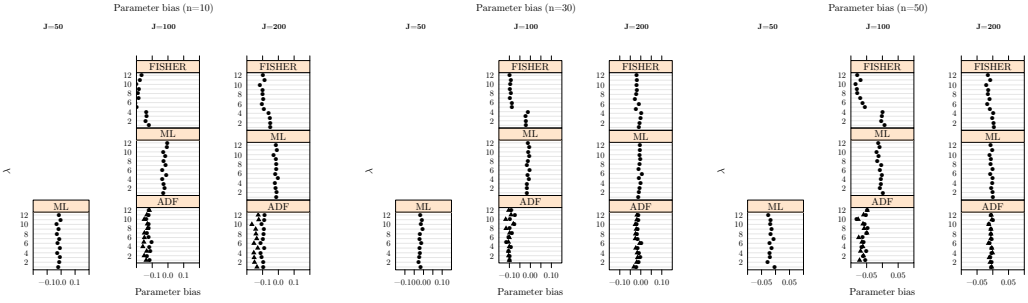
(a) $ICC = .50$ $n = 10$ (b) $ICC = .50$ $n = 30$ (c) $ICC = .50$ $n = 50$



(d) $ICC = .26$ $n = 10$ (e) $ICC = .26$ $n = 30$ (f) $ICC = .26$ $n = 50$



(g) $ICC = .10$ $n = 10$ (h) $ICC = .10$ $n = 30$ (i) $ICC = .10$ $n = 50$

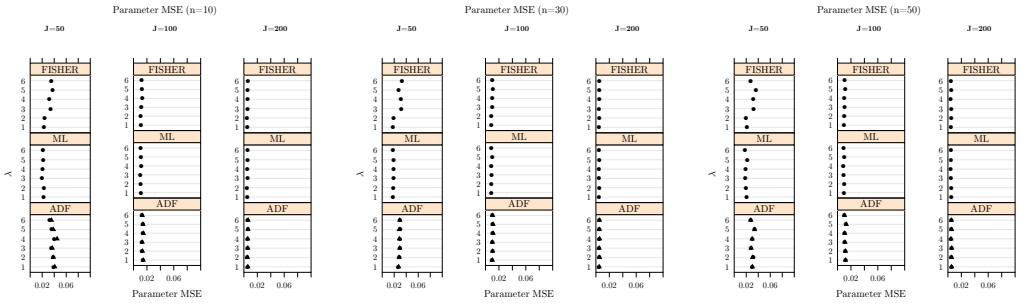


(j) $ICC = .05$ $n = 10$ (k) $ICC = .05$ $n = 30$ (l) $ICC = .05$ $n = 50$

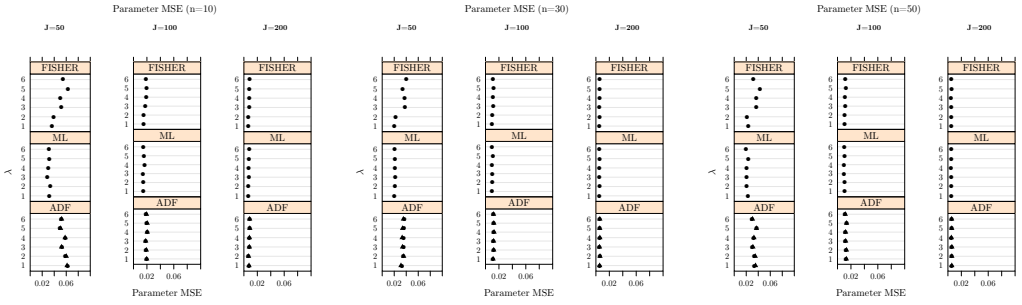
Legend: ▲ =Bootstrap, • =GEE

panel displays MSE based on the ML estimator. The use of the bootstrap does not influence the ML parameter estimates. The third panel displays MSE based on ADF estimation using $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$. The GEE-based estimates are represented by black circles (●), and the cluster bootstrap based estimates are represented by black triangles (▲). Figure 6.3 displays only information for the small models ($df = 9$). Figure 6.4 displays only information for the large models ($df = 54$). Lower MSE would imply that the dots in each panel should be towards the left, and the closer the plotted points are to the vertical axis, the lower the MSE. A consistent estimator would result in plots that show a trend of decreasing MSE, as sample sizes increase. For all of the estimators, regardless of within group sample size or ICC, the MSE decreases as the number of groups increases, showing that both ADF estimators, and the ML estimator are consistent when used in conjunction with the segregating method. The mean square error is much larger in conditions with low level-2 sample sizes, small within-group sample sizes, and low ICCs (for example, the $ICC = .05$, $n = 10$ condition in Figure 6.3j) than it is in conditions with large level 2 sample sizes, large within-group sample sizes, and high ICCs (for example, the $ICC = .50$, $n = 50$ condition in Figure 6.3c). Additionally, with small level-2 sample sizes, small within-group sample sizes, and low ICCs, the differences between the GEE and bootstrap-based estimators is more pronounced. Specifically, the GEE- based estimator has smaller mean square error than the bootstrap-based estimator. This pattern is also true for larger models (Figure 6.4). Of note is the fact that, for most conditions, the ML estimator has smaller mean square error than either ADF approach. These differences disappear as the number of groups increase.

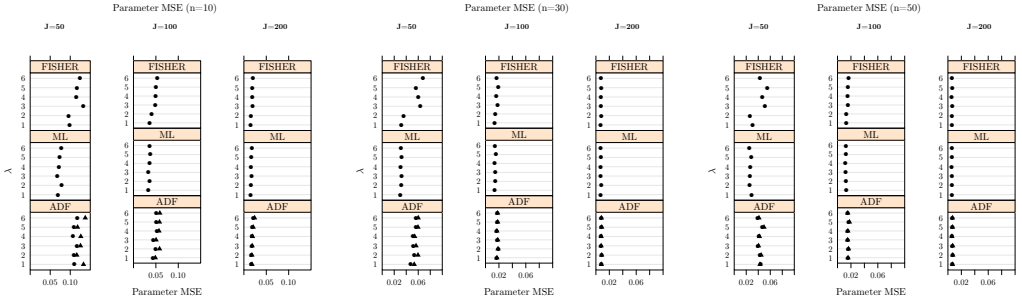
Figure 6.3: Mean square error by estimator: $df = 9$



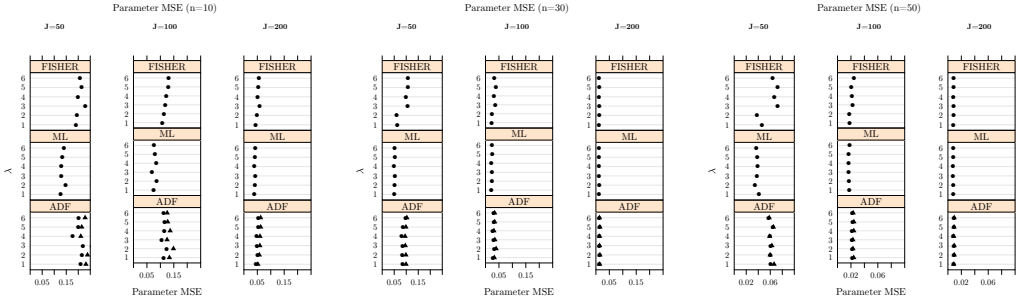
(a) $ICC = .50$ $n = 10$ (b) $ICC = .50$ $n = 30$ (c) $ICC = .50$ $n = 50$



(d) $ICC = .26$ $n = 10$ (e) $ICC = .26$ $n = 30$ (f) $ICC = .26$ $n = 50$



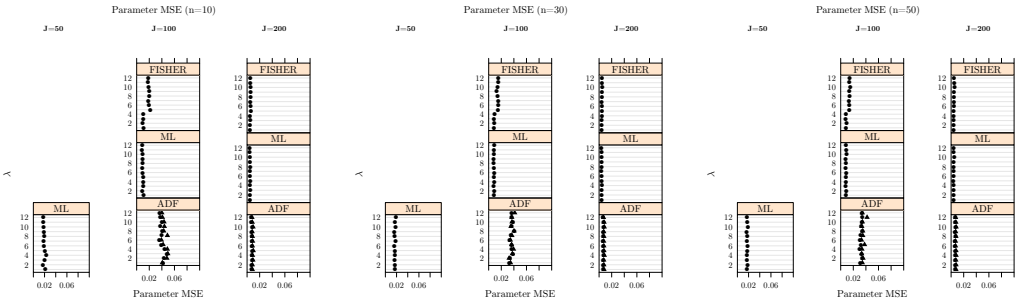
(g) $ICC = .10$ $n = 10$ (h) $ICC = .10$ $n = 30$ (i) $ICC = .10$ $n = 50$



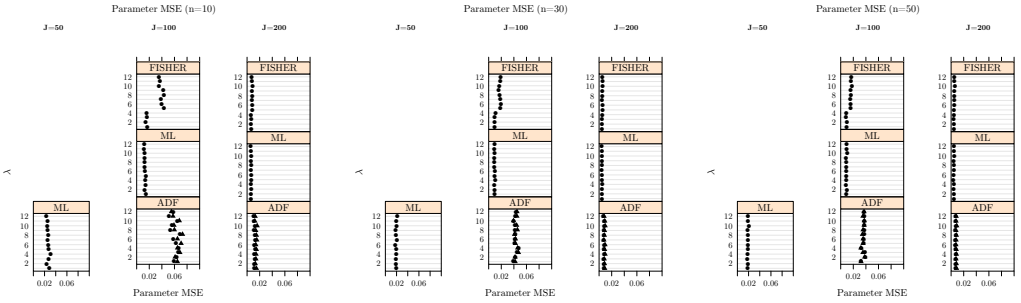
(j) $ICC = .05$ $n = 10$ (k) $ICC = .05$ $n = 30$ (l) $ICC = .05$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

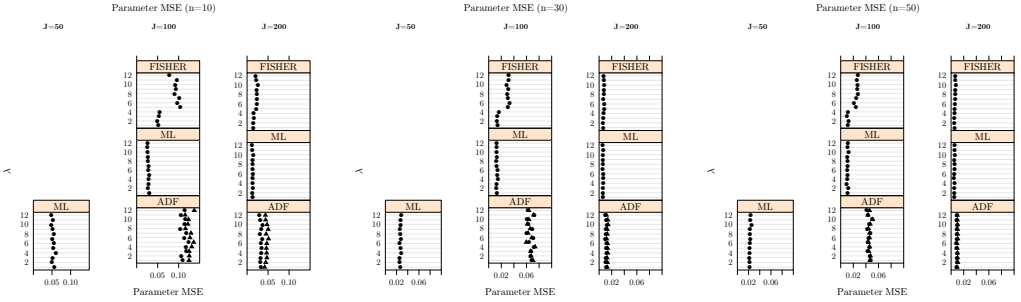
Figure 6.4: Mean square error by estimator: $df = 54$



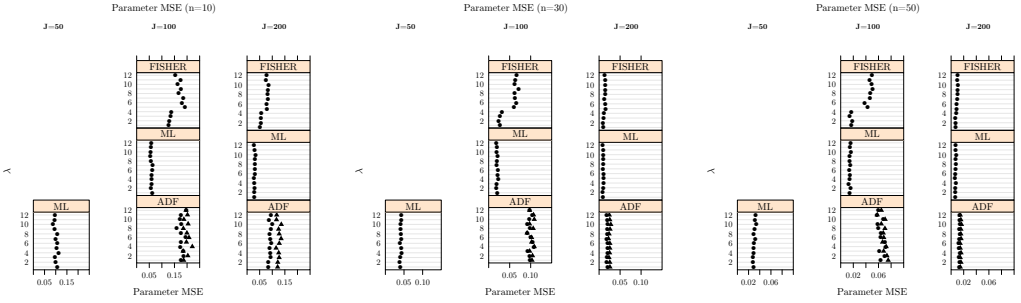
(a) $ICC = .50$ $n = 10$ (b) $ICC = .50$ $n = 30$ (c) $ICC = .50$ $n = 50$



(d) $ICC = .26$ $n = 10$ (e) $ICC = .26$ $n = 30$ (f) $ICC = .26$ $n = 50$



(g) $ICC = .10$ $n = 10$ (h) $ICC = .10$ $n = 30$ (i) $ICC = .10$ $n = 50$



(j) $ICC = .05$ $n = 10$ (k) $ICC = .05$ $n = 30$ (l) $ICC = .05$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

6.2.3 Are the standard error estimates consistent using the segregating method?

Since standard errors for model parameters based on Γ_{FISHER} are correct under normality (e.g. Yuan & Hayashi, 2006), the consistency of the standard error estimates based on the ML estimator, the robust estimator, and the ADF-theory estimators using $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$ were assessed using D^2 as given in Equation 5.5 . If these estimated standard errors are consistent, D^2 should approach zero (e.g. Yuan & Hayashi, 2006).

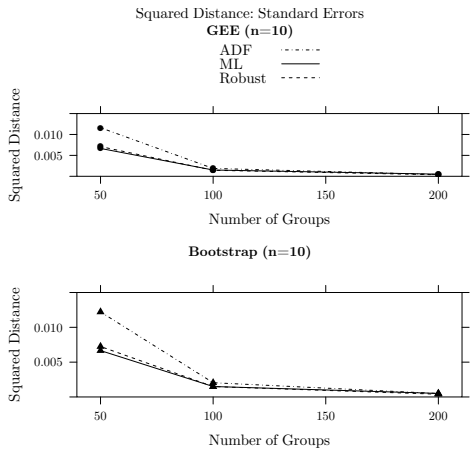
Figures 6.5- 6.12 display results for all model conditions. Information about ML estimated standard errors is included in both panels for reference, and shown as a solid line. The use of GEE or bootstrap based approaches does not influence the ML estimated standard errors. The GEE-based estimates are represented by black circles (•), and the cluster bootstrap based estimates are represented by black triangles (▲).

From these plots, it can be seen that standard errors based on ML do not show good convergence properties at low ICC conditions, particularly with small sample sizes. For example, in Figures 6.7a and 6.8a, the D^2 value is much higher for the ML standard errors at $J = 200$ than for the ADF or robust standard errors. On the other hand, the robust standard errors and the ADF standard errors show good converge properties—D^2 decreases as level 2 sample sizes increase for all ICC and within-group sample size conditions. For $J = 200$, D^2 is essentially zero for the robust and ADF estimators. With sufficient sample sizes ($J = 200$) there is almost no difference between the standard errors based on ADF estimators using $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$. These results suggest that the cluster bootstrap yields consistent standard error estimates, and specifically that for sufficient sample sizes, the cluster bootstrap performs as well as GEE.

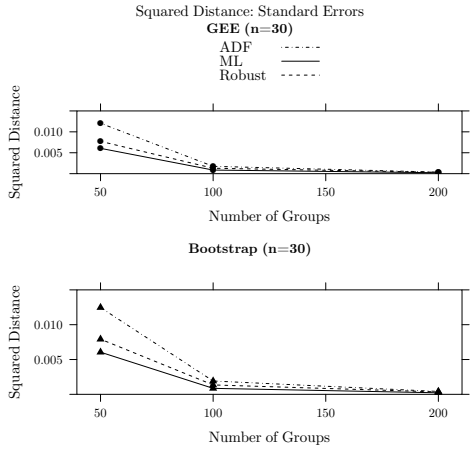
Similar patterns hold for the larger models ($df = 54$), which are displayed in Figures 6.9-6.12. Specifically, ML estimates of standard errors do not show good

Figure 6.5: D^2 plots: standard errors, $df = 9$ $ICC = .50$

(a)



(b)



(c)

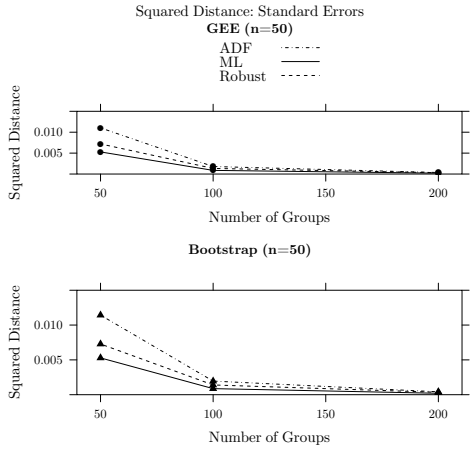
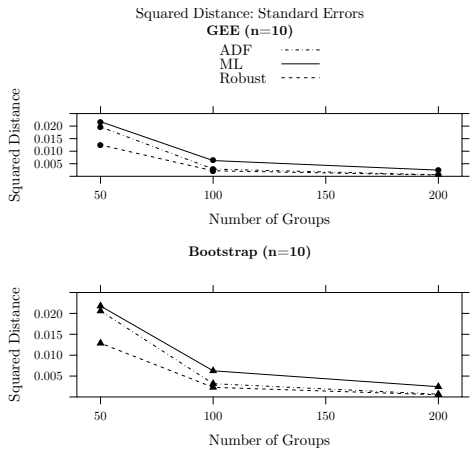
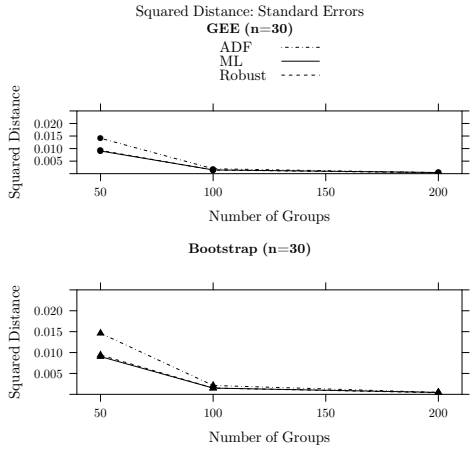


Figure 6.6: D^2 plots: standard errors, $df = 9$ $ICC = .26$

(a)



(b)



(c)

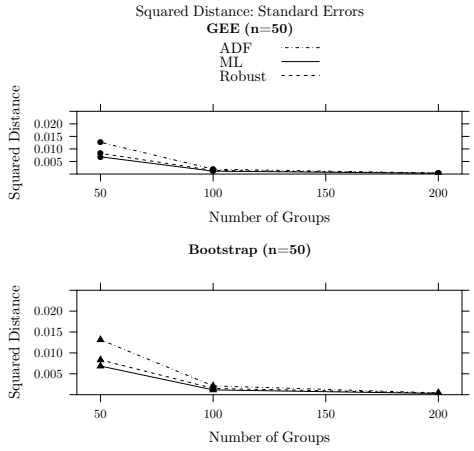
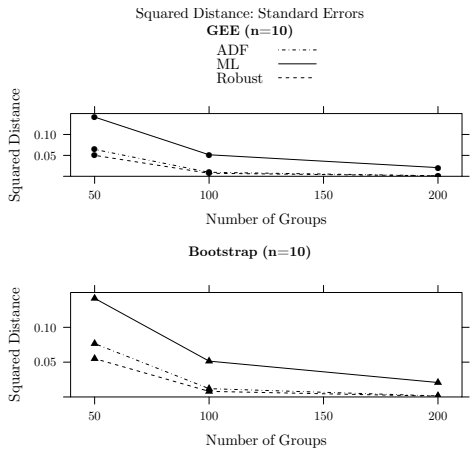
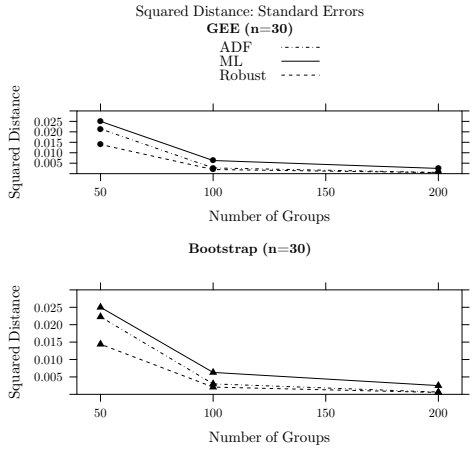


Figure 6.7: D^2 plots: standard errors, $df = 9$ $ICC = .10$

(a)



(b)



(c)

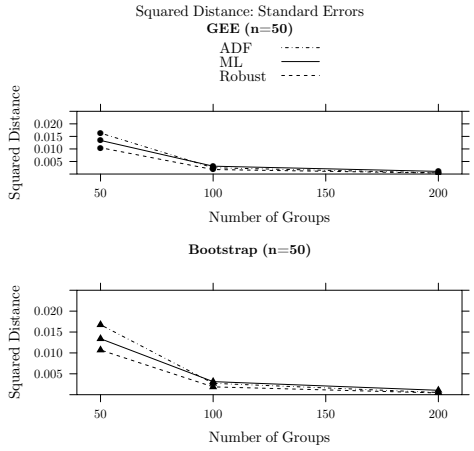
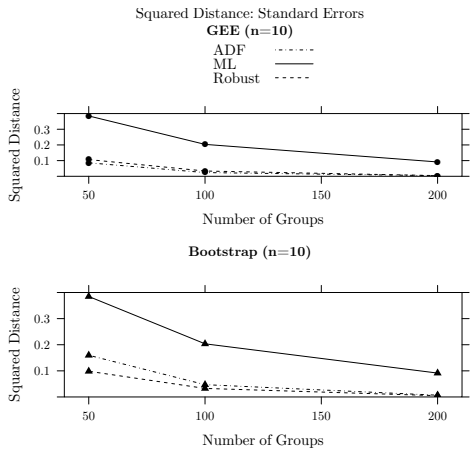
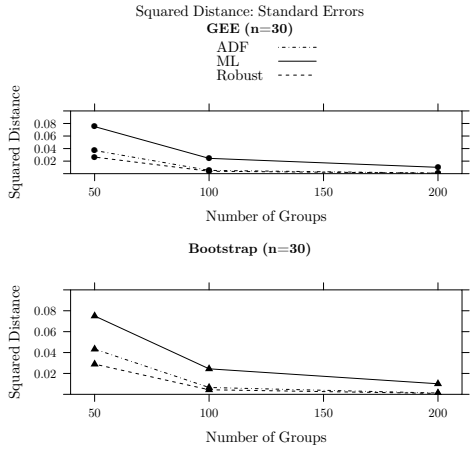


Figure 6.8: D^2 plots: standard errors, $df = 9$ $ICC = .05$

(a)



(b)



(c)

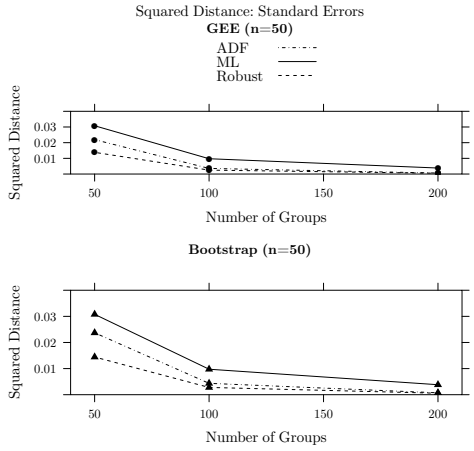
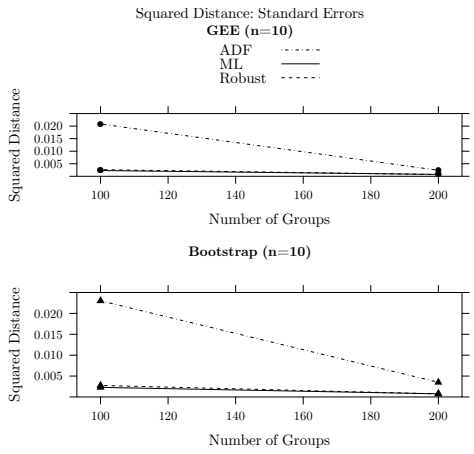
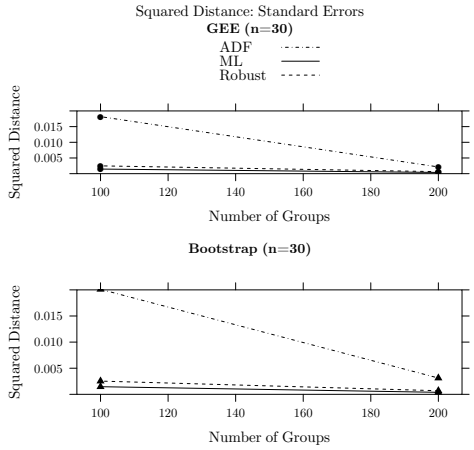


Figure 6.9: D^2 plots: standard errors, $df = 54$ $ICC = .50$

(a)



(b)



(c)

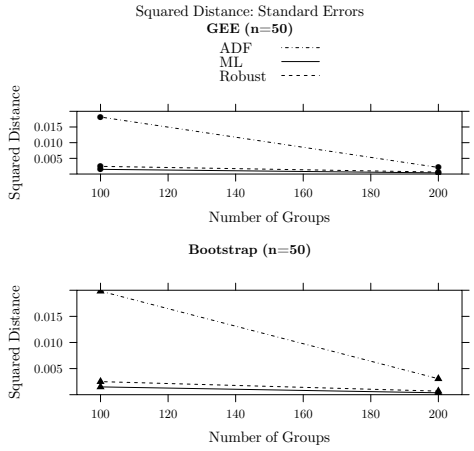
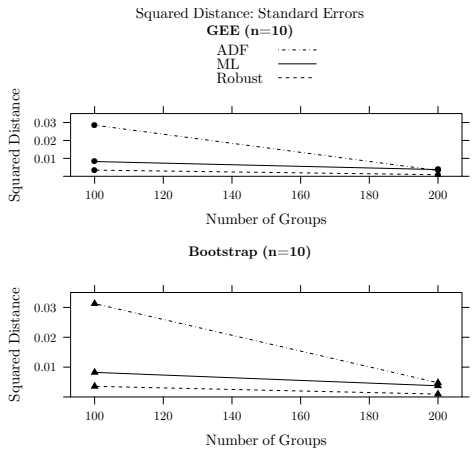
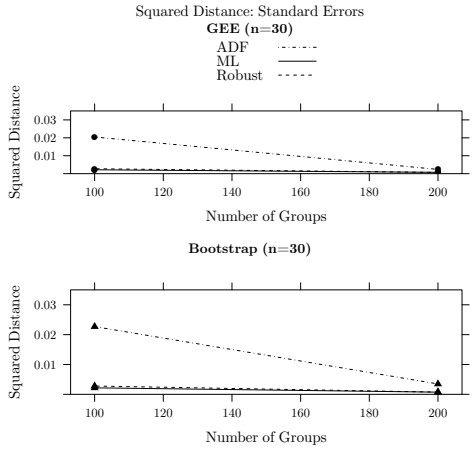


Figure 6.10: D^2 plots: standard errors, $df = 54$ $ICC = .26$

(a)



(b)



(c)

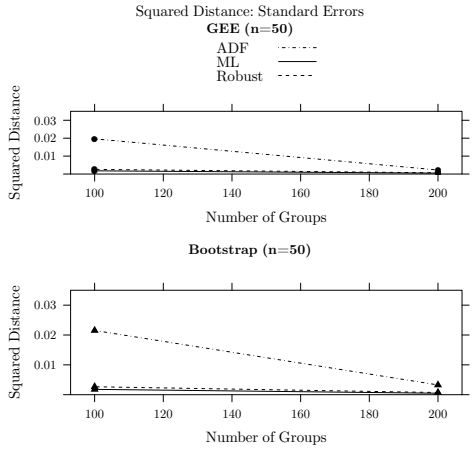
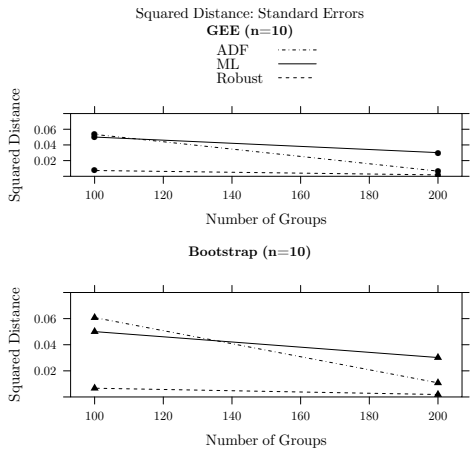
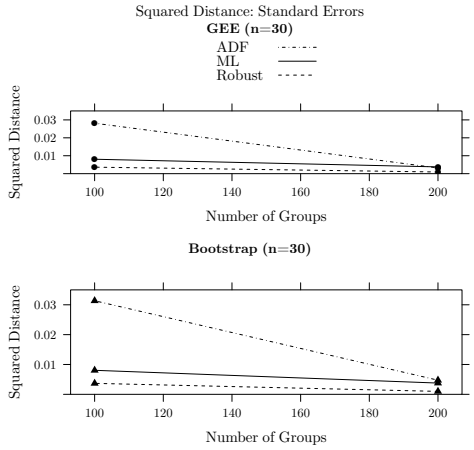


Figure 6.11: D^2 plots: standard errors, $df = 54$ $ICC = .10$

(a)



(b)



(c)

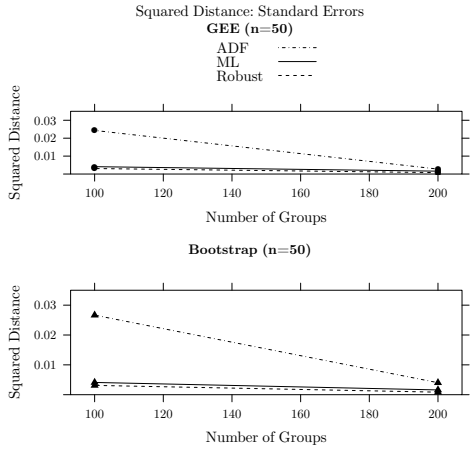
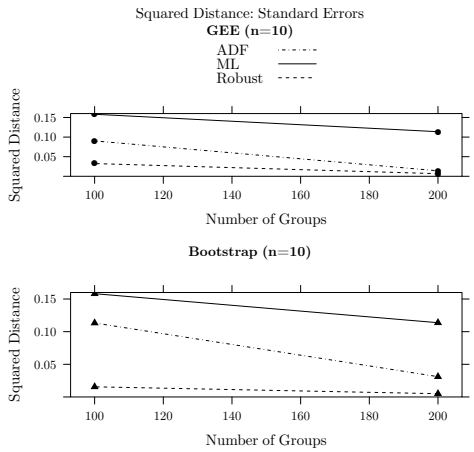
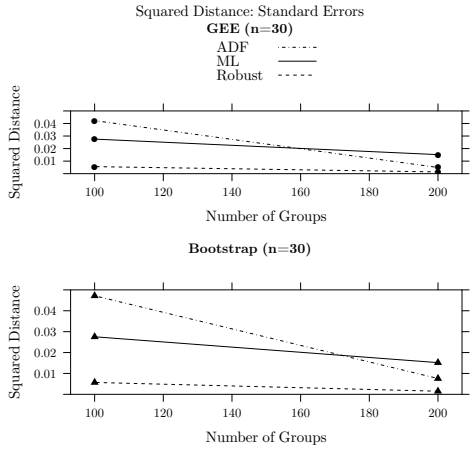


Figure 6.12: D^2 plots: standard errors, $df = 54$ $ICC = .05$

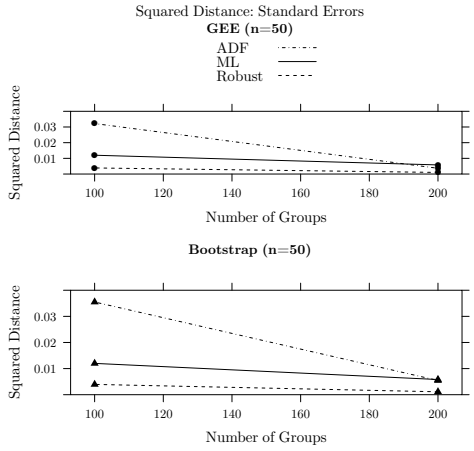
(a)



(b)



(c)



convergence for low ICCs and small within-group sample sizes. This can be seen clearly in Figure 6.11a and 6.12a, where the plot of D^2 for the ML estimator is nearly a horizontal line. Robust and ADF standard errors based either on $\hat{\Gamma}_{GEE}$ or $\hat{\Gamma}_{BOOT}$ do show good convergence, and for all combinations of ICC and within group sample size, as the level-2 sample size increases, D^2 decreases, and is almost negligible at $J = 200$. The robust standard errors show noticeably lower D^2 than the ADF standard errors at all sample sizes.

6.2.4 Test statistic distributions

6.2.4.1 Type I error rates

Tables 6.3 and 6.4 present the empirical type I error rates for all six test statistics, as well as T_{FISHER}, the ADF test statistic that uses Γ_{FISHER}. Under normality, Γ_{FISHER} is correctly specified, and so T_{FISHER} will have the correct asymptotic distribution as a central chi square variate. The five test statistics that use either $\hat{\Gamma}_{GEE}$ or $\hat{\Gamma}_{BOOT}$, namely T_{ADF}, T_{CADF}, T_{RML}, T_{RADF} and T_{CRADF}, are presented side by side for direct comparison of the GEE and cluster bootstrap approaches. Because it is expected that the empirical error rates will differ somewhat from the nominal rate, an acceptable empirical error rate is taken as one that falls in the interval [.028, .079], the estimated 2-sided 99% adjusted Wald confidence interval (Agresti & Coull, 1998). Empirical rejection rates in this interval are shown in bold.

Table 6.3: Empirical Type I error rates, $df = 9$.

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
		(a) $J = 200$					
T_{FISHER}	$ICC = .50$	**0.054**		**0.036**		**0.036**	
	$ICC = .26$	**0.058**		**0.038**		**0.034**	

Table 6.3 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
	$ICC = .10$	**0.036**		**0.048**		**0.034**	
	$ICC = .05$	0.018		**0.034**		0.026	
T_{ML}	$ICC = .50$	0.198		0.088		**0.074**	
	$ICC = .26$	0.492		0.186		0.104	
	$ICC = .10$	0.962		0.544		0.336	
	$ICC = .05$	1.000		0.872		0.668	
T_{RML}	$ICC = .50$	**0.074**	**0.076**	**0.064**	**0.066**	**0.062**	**0.062**
	$ICC = .26$	**0.058**	**0.060**	**0.052**	**0.058**	**0.056**	**0.060**
	$ICC = .10$	0.114	0.140	**0.070**	**0.072**	**0.064**	**0.068**
	$ICC = .05$	0.130	0.270	0.102	0.110	0.088	0.090
T_{ADF}	$ICC = .50$	0.090	0.094	0.086	0.096	0.078	0.104
	$ICC = .26$	**0.076**	0.092	**0.072**	0.094	0.084	0.096
	$ICC = .10$	**0.066**	0.110	**0.072**	0.090	0.082	0.088
	$ICC = .05$	0.012	0.128	**0.070**	0.106	**0.072**	0.090
T_{CADF}	$ICC = .50$	**0.064**	**0.068**	**0.048**	**0.058**	**0.052**	**0.066**
	$ICC = .26$	**0.050**	**0.064**	**0.044**	**0.042**	**0.052**	**0.058**
	$ICC = .10$	**0.038**	**0.072**	**0.046**	**0.048**	**0.052**	**0.062**
	$ICC = .05$	0.002	0.094	**0.044**	**0.072**	**0.052**	**0.058**
T_{RADF}	$ICC = .50$	0.096	0.094	0.088	0.096	0.078	0.104
	$ICC = .26$	**0.076**	0.092	**0.072**	0.094	0.086	0.096
	$ICC = .10$	**0.070**	0.118	**0.076**	0.090	0.082	0.098
	$ICC = .05$	0.026	0.152	**0.072**	0.110	**0.076**	0.098
T_{CRADF}	$ICC = .50$	**0.064**	**0.068**	**0.050**	**0.058**	**0.054**	**0.068**
	$ICC = .26$	**0.052**	**0.064**	**0.044**	**0.044**	**0.054**	**0.060**
	$ICC = .10$	**0.038**	**0.078**	**0.050**	**0.056**	**0.052**	**0.066**
	$ICC = .05$	0.008	0.122	**0.048**	**0.078**	**0.052**	**0.054**
		(b) $J = 100$					
T_{FISHER}	$ICC = .50$	**0.044**		0.018		**0.040**	
	$ICC = .26$	**0.038**		**0.028**		**0.050**	
	$ICC = .10$	0.018		**0.034**		**0.050**	

Table 6.3 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
	$ICC = .05$	0.004		0.020		**0.046**	
T_{ML}	$ICC = .50$	0.190		0.092		0.102	
	$ICC = .26$	0.514		0.180		0.158	
	$ICC = .10$	0.966		0.592		0.382	
	$ICC = .05$	1.000		0.924		0.690	
T_{RML}	$ICC = .50$	**0.072**	**0.076**	**0.050**	**0.058**	0.086	0.092
	$ICC = .26$	0.098	0.102	**0.064**	**0.066**	**0.074**	**0.076**
	$ICC = .10$	0.198	0.266	0.086	0.094	0.086	0.096
	$ICC = .05$	0.096	0.304	0.124	0.154	0.128	0.138
T_{ADF}	$ICC = .50$	0.120	0.138	0.116	0.122	0.142	0.154
	$ICC = .26$	0.124	0.128	0.112	0.114	0.138	0.152
	$ICC = .10$	**0.046**	0.170	0.102	0.116	0.132	0.146
	$ICC = .05$	0.008	0.082	**0.074**	0.140	0.140	0.160
T_{CADF}	$ICC = .50$	**0.058**	**0.060**	**0.056**	**0.052**	**0.068**	0.078
	$ICC = .26$	**0.040**	**0.056**	**0.044**	**0.050**	**0.07**	**0.068**
	$ICC = .10$	0.014	0.090	**0.040**	**0.050**	**0.066**	**0.074**
	$ICC = .05$	0.002	0.026	0.024	**0.074**	**0.064**	0.092
T_{RADF}	$ICC = .50$	0.122	0.146	0.118	0.130	0.152	0.154
	$ICC = .26$	0.134	0.138	0.122	0.130	0.150	0.160
	$ICC = .10$	**0.070**	0.194	0.108	0.128	0.140	0.154
	$ICC = .05$	0.020	0.148	0.086	0.178	0.150	0.170
T_{CRADF}	$ICC = .50$	**0.064**	**0.070**	**0.054**	**0.060**	**0.070**	0.080
	$ICC = .26$	**0.048**	**0.068**	**0.048**	**0.056**	**0.076**	**0.078**
	$ICC = .10$	0.024	0.112	**0.044**	**0.056**	**0.076**	0.084
	$ICC = .05$	0.006	**0.058**	**0.034**	0.086	0.078	0.110
		(c) $J = 50$					
T_{FISHER}	$ICC = .50$	0.022		0.012		**0.044**	
	$ICC = .26$	0.028		**0.040**		**0.034**	
	$ICC = .10$	0.006		0.010		**0.040**	
	$ICC = .05$	0.000		0.012		**0.030**	

Table 6.3 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
T_{ML}	$ICC = .50$	0.258		0.112		0.114	
	$ICC = .26$	0.582		0.214		0.178	
	$ICC = .10$	0.978		0.656		0.436	
	$ICC = .05$	0.998		0.924		0.778	
T_{RML}	$ICC = .50$	0.096	0.112	0.084	0.100	0.102	0.116
	$ICC = .26$	0.164	0.182	0.090	0.104	0.114	0.118
	$ICC = .10$	0.222	0.438	0.170	0.204	0.132	0.146
	$ICC = .05$	0.096	0.240	0.124	0.356	0.128	0.304
T_{ADF}	$ICC = .50$	0.154	0.154	0.168	0.174	0.178	0.204
	$ICC = .26$	0.132	0.162	0.160	0.166	0.184	0.190
	$ICC = .10$	**0.042**	0.172	0.154	0.200	0.188	0.202
	$ICC = .05$	0.000	0.024	0.094	0.218	0.180	0.082
T_{CADF}	$ICC = .50$	0.028	0.028	0.016	0.016	0.022	**0.032**
	$ICC = .26$	**0.028**	**0.038**	0.026	**0.036**	0.026	**0.034**
	$ICC = .10$	0.000	0.020	0.016	**0.044**	0.024	**0.042**
	$ICC = .05$	0.000	0.000	0.004	**0.030**	0.026	0.026
T_{RADF}	$ICC = .50$	0.198	0.210	0.204	0.204	0.224	0.238
	$ICC = .26$	0.188	0.238	0.198	0.220	0.226	0.246
	$ICC = .10$	0.082	0.232	0.206	0.268	0.25	0.262
	$ICC = .05$	0.006	0.088	0.174	0.304	0.238	0.148
T_{CRADF}	$ICC = .50$	**0.054**	**0.054**	**0.036**	**0.042**	**0.066**	**0.074**
	$ICC = .26$	**0.038**	**0.062**	**0.054**	**0.054**	**0.056**	**0.064**
	$ICC = .10$	0.004	**0.036**	**0.046**	0.084	**0.044**	**0.050**
	$ICC = .05$	0.000	0.002	0.018	**0.074**	**0.040**	**0.058**

As anticipated based on previous research (Schweig, 2014), because of the clustering effect, as either ICC or within group sample size decrease Type I error rates increase for T_{ML}. T_{ML} only has acceptable Type I error rates with small models, a larger

number of groups, 50 individuals per group and an ICC of .50. With $ICC = .05$, $n = 10$, the correct model is rejected 100% of the time based on T_{ML}, even with a large number of groups. With larger models (Table 6.4), the empirical Type I error rates are consistently too high for T_{ML} across all model conditions, and never get below .132 ($ICC = .50$, $n = 50$, $J = 200$). Of the six test statistics that use either $\hat{\Gamma}_{GEE}$ or $\hat{\Gamma}_{BOOT}$, only the corrected test statistics, T_{CADF} and T_{CRADF} show good rejection rates over a wide range of conditions. Specifically, both T_{CADF} and T_{CRADF} generally have empirical Type I error rates in the interval [.028,.076]. One exception to this is when ICCs are low ($ICC = .05$) and within group sample sizes are small ($n = 10$). There, the Type I error rates are too low for the GEE-based estimator (.002 and .008, for T_{CADF} and T_{CRADF}, respectively) and too high for the bootstrap-based estimator (.094 and .122, for T_{CADF} and T_{CRADF}, respectively). The rescaled test statistic T_{RML} performs better than the conventional likelihood ratio test statistic T_{ML}, and for small models, relatively large sample sizes and high ICCs, T_{RML} has acceptable rejection rates. For example, in Table 6.3a, for $J = 200$, T_{RML} has acceptable rejection rates at $ICC = .26$ and $ICC = .50$ for all within-group sample sizes. However, when the full range of simulation conditions are considered, it becomes clear that T_{RML} cannot adequately control Type I errors when group sizes are small or when ICCs are low. For example, in Table 6.3c, T_{RML} shows rejection rates above .40 when ICCs are low ($ICC = .10$) and within group sample sizes are small ($n = 10$).

The rejection rates in Table 6.3 also demonstrate that, for sufficient sample sizes, the cluster bootstrap based test statistics give inferences that are consistent with those based on GEE estimation. The Type I error rates are generally comparable for the two approaches for all models in all conditions. In particular, if the Type I error rate for the GEE-based test statistics are in the interval [.028,.076], nearly 80% of the Bootstrap-based test statistics are in that interval, as well.

Table 6.4 shows empirical Type I errors for the larger models ($df = 54$). When

the model is large and the number of groups is small, T_{RADF} and T_{ADF} reject the correct model nearly 100% of the time. Even with 200 groups (Table 6.4a) the empirical Type I error rates approach 90%. This is consistent with past results showing that ADF test statistics converge slowly to the appropriate distribution (e.g., Curran et al., 1996; Hu et al., 1992; B. O. Muthén & Kaplan, 1985, 1992; Yuan & Bentler, 2003; Bentler & Yuan, 1999; Powell & Schafer, 2001). T_{RML} is also unable to control Type I errors when the model is large. Even with 200 groups (Table 6.4a), Type I error rates are as high as .600 (with $ICC = .10$ and $n = 10$). Whereas the rejection rates in Table 6.3 demonstrated that the cluster bootstrap based test statistics give inferences that are consistent with those based on GEE estimation, the results in Table 6.4 show that, for larger models, this is not the case. Specifically, the Type I error rates for T_{ADF} and T_{RADF} are systematically higher for the bootstrap based approach. Even with $J = 200$ (Table 6.4a), the Type I error rates for T_{ADF} and T_{RADF} are as much as 30% larger based on the bootstrap based approach.

The corrected test statistics, T_{CRADF} and T_{CADF} based on the GEE estimator show good empirical rejection rates for sufficiently large sample sizes. Specifically, empirical rejection rates are in the interval [.028,.076] for nearly all ICC and within-group sample size combination when $J = 200$ (Table 6.4a). However, when the number of groups is small relative to the size of the model (for example, in Table 6.4b), the multilevel version of T_{CRADF} performs similarly to the conventional version (Bentler & Yuan, 1999). Specifically, the Type I error rate is too low—in many conditions, no models are rejected by either T_{CRADF} or T_{CADF}. On the other hand, the bootstrap-based versions of T_{CRADF} and T_{CADF} are unable to appropriately correct the statistics for larger sample sizes (Table 6.4a), and the Type I error rates for these bootstrap-based statistics are systematically too high. This is a direct result of the performance of the bootstrap-based T_{RADF} and T_{ADF} test statistics, which are systematically larger than their GEE-based counterparts.

For smaller level-2 sample sizes, the bootstrap-based versions of T_{CRADF} and T_{CADF} behave like the GEE-based versions, and the observed Type I error rates are too low.

Table 6.4: Empirical Type I error rates, $df = 54$.

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
		(a) $J = 200$					
T_{FISHER}	$ICC = .50$	**0.040**		**0.040**		**0.056**	
	$ICC = .26$	**0.044**		**0.040**		**0.048**	
	$ICC = .10$	0.028		**0.038**		**0.05**	
	$ICC = .05$	0.006		**0.038**		**0.054**	
T_{ML}	$ICC = .50$	0.508		0.192		0.132	
	$ICC = .26$	0.962		0.432		0.264	
	$ICC = .10$	1.000		0.976		0.840	
	$ICC = .05$	1.000		1.000		0.996	
T_{RML}	$ICC = .50$	0.082	0.082	0.088	0.090	0.084	0.086
	$ICC = .26$	0.172	0.188	0.102	0.110	0.092	0.092
	$ICC = .10$	0.492	0.600	0.184	0.186	0.122	0.130
	$ICC = .05$	0.540	0.916	0.354	0.406	0.224	0.232
T_{ADF}	$ICC = .50$	0.638	0.782	0.648	0.772	0.654	0.796
	$ICC = .26$	0.638	0.796	0.628	0.806	0.634	0.792
	$ICC = .10$	0.548	0.876	0.622	0.790	0.658	0.804
	$ICC = .05$	0.156	0.870	0.606	0.796	0.646	0.798
T_{CADF}	$ICC = .50$	**0.048**	0.122	**0.048**	0.134	**0.042**	0.128
	$ICC = .26$	**0.054**	0.138	**0.05**	0.126	**0.052**	0.134
	$ICC = .10$	0.028	0.258	**0.044**	0.132	**0.05**	0.116
	$ICC = .05$	0.000	0.132	**0.03**	0.216	**0.046**	0.156
T_{RADF}	$ICC = .50$	0.648	0.800	0.662	0.794	0.668	0.802
	$ICC = .26$	0.660	0.802	0.656	0.824	0.66	0.804
	$ICC = .10$	0.612	0.902	0.642	0.820	0.684	0.818
	$ICC = .05$	0.312	0.948	0.658	0.832	0.668	0.818
T_{CRADF}	$ICC = .50$	**0.054**	0.150	**0.054**	0.150	**0.052**	0.160

Table 6.4 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
	$ICC = .26$	**0.068**	0.176	**0.058**	0.152	**0.056**	0.164
	$ICC = .10$	**0.04**	0.350	**0.054**	0.162	**0.058**	0.140
	$ICC = .05$	0.004	0.312	**0.048**	0.264	**0.064**	0.196
		(b) $J = 100$					
T_{FISHER}	$ICC = .50$	0.020		**0.044**		**0.038**	
	$ICC = .26$	0.016		**0.048**		**0.036**	
	$ICC = .10$	0.004		**0.032**		**0.040**	
	$ICC = .05$	0.000		0.012		0.026	
T_{ML}	$ICC = .50$	0.63		0.272		0.172	
	$ICC = .26$	0.982		0.564		0.350	
	$ICC = .10$	1.000		0.992		0.894	
	$ICC = .05$	1.000		1.000		1.000	
T_{RML}	$ICC = .50$	0.170	0.176	0.152	0.160	0.108	0.128
	$ICC = .26$	0.370	0.396	0.194	0.212	0.136	0.158
	$ICC = .10$	0.768	0.904	0.37	0.402	0.234	0.248
	$ICC = .05$	0.464	0.964	0.658	0.762	0.45	0.506
T_{ADF}	$ICC = .50$	0.978	0.984	0.986	0.998	0.972	0.990
	$ICC = .26$	0.972	0.988	0.984	0.990	0.974	0.998
	$ICC = .10$	0.928	0.992	0.976	0.994	0.974	0.988
	$ICC = .05$	0.468	0.916	0.964	0.992	0.972	0.992
T_{CADF}	$ICC = .50$	0.002	0.006	0.000	0.002	0.000	0.000
	$ICC = .26$	0.000	0.006	0.002	0.000	0	0.002
	$ICC = .10$	0.000	0.002	0	0.002	0.002	0.006
	$ICC = .05$	0.000	0.000	0.002	0.004	0	0.010
T_{RADF}	$ICC = .50$	0.988	0.996	0.99	1.000	0.986	0.992
	$ICC = .26$	0.986	0.992	0.986	0.998	0.984	0.998
	$ICC = .10$	0.974	0.992	0.988	0.994	0.982	0.990
	$ICC = .05$	0.682	0.978	0.98	0.996	0.982	0.992
T_{CRADF}	$ICC = .50$	0.008	0.018	0.006	0.006	0.006	0.010
	$ICC = .26$	0.012	0.028	0.008	0.016	0.006	0.010

Table 6.4 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
	$ICC = .10$	0.000	0.010	0.01	0.024	0.006	0.016
	$ICC = .05$	0.000	0.000	0.004	0.016	0.006	**0.034**
				(b) $J = 50$			
T_{ML}	$ICC = .50$	0.788		0.380		0.288	
	$ICC = .26$	0.998		0.732		0.53	
	$ICC = .10$	1.000		1.000		0.966	
	$ICC = .05$	1.000		1.000		1.000	
T_{RML}	$ICC = .50$	0.342	0.386	0.244	0.266	0.22	0.262
	$ICC = .26$	0.644	0.722	0.358	0.394	0.284	0.326
	$ICC = .10$	0.864	0.984	0.748	0.800	0.542	0.598
	$ICC = .05$	0.226	0.926	0.914	0.99	0.834	0.902
T_{RADF}	$ICC = .50$		1.000		1.000		1.000
	$ICC = .26$		1.000		1.000		1.000
	$ICC = .10$		0.998		1.000		1.000
	$ICC = .05$		0.980		1.000		1.000
T_{CRADF}	$ICC = .50$		0.000		0.000		0.000
	$ICC = .26$		0.000		0.000		0.000
	$ICC = .10$		0.000		0.000		0.000
	$ICC = .05$		0.000		0.000		0.000

When the sample size is small ($J = 50$) and the model is large ($df = 54$), the GEE based test statistics T_{RADF} and T_{CRADF} are not computable because $[\dot{\sigma}_c(\hat{\theta})^T \hat{\Gamma}_{GEE} \dot{\sigma}_c(\hat{\theta})]^{-1}$ (3.19) is not invertible. Interestingly, the bootstrap-based versions of these test statistics are computable. However, the performance of these statistics is quite poor. The bootstrap-based T_{RADF} rejects nearly every model in every condition (Table 6.4c). On the other hand, the corrected test statistic T_{CRADF} failed to reject a single model. In fact, for small sample sizes and large

models, as presented in Table 6.4c, no test statistics show adequate Type I error rates.

6.2.4.2 Test statistic means and standard deviations

A test statistic is called well-behaved if its empirical distribution is similar to its theoretical distribution. For the small models ($df = 9$), the theoretical chi-square distribution has a mean of 9 and a standard deviation of $\sqrt{18} \approx 4.24$. Likewise, for the large models ($df = 54$), the central chi-square distribution has a mean of 54 and a standard deviation of $\sqrt{108} \approx 10.29$. Thus, the empirical means and standard deviations of the estimated chi-square test statistics should be close to these values in order for these statistics to be considered well-behaved. However, while the asymptotic properties of test statistic distributions are well known (with the exception of T_{RML}, which is only theorized to have the correct mean), the small sample properties are generally unknown (Bentler & Chou, 1987; Tanaka, 1987). Thus, it is important to consider not only the empirical means and standard deviations, but also whether or not these values converge to their theoretical values as sample sizes increase.

Table 6.5: Test statistic means and standard deviations, $df = 9$.

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
		(a) $J = 200$					
T_{FISHER}	$ICC = .50$	8.80 (4.40)		8.75 (4.13)		8.81 (4.15)	
	$ICC = .26$	8.67 (4.34)		8.73 (4.11)		8.81 (4.07)	
	$ICC = .10$	8.28 (4.05)		8.68 (4.05)		8.77 (3.92)	
	$ICC = .05$	7.21 (3.33)		8.58 (3.95)		8.70 (3.83)	
T_{ML}	$ICC = .50$	12.34 (6.09)		10.10 (4.72)		9.76 (4.52)	
	$ICC = .26$	18.32 (9.54)		11.71 (5.55)		10.74 (4.98)	
	$ICC = .10$	63.00 (50.90)		19.72 (10.07)		15.13 (7.39)	
	$ICC = .05$	193.73 (120.24)	39.36 (26.93)		24.25 (14.80)		
T_{RML}	$ICC = .50$	9.42 (4.61)	9.47 (4.61)	9.25 (4.29)	9.31 (4.33)	9.30 (4.31)	9.35 (4.35)
	$ICC = .26$	9.42 (4.83)	9.48 (4.88)	9.27 (4.38)	9.31 (4.39)	9.35 (4.34)	9.39 (4.37)
	$ICC = .10$	10.17 (7.29)	10.94 (8.61)	9.44 (4.73)	9.50 (4.77)	9.51 (4.61)	9.57 (4.64)
	$ICC = .05$	9.91 (6.39)	13.62 (9.03)	10.01 (6.26)	10.27 (6.64)	9.87 (5.75)	9.96 (5.83)
T_{ADF}	$ICC = .50$	9.92 (4.99)	10.16 (5.12)	9.78 (4.70)	10.02 (4.91)	9.83 (4.75)	10.08 (4.84)
	$ICC = .26$	9.72 (4.82)	9.94 (5.00)	9.73 (4.62)	9.96 (4.76)	9.86 (4.80)	10.07 (4.94)
	$ICC = .10$	9.24 (4.34)	10.21 (5.60)	9.68 (4.57)	9.94 (4.76)	9.89 (4.81)	10.17 (4.96)
	$ICC = .05$	7.96 (3.42)	10.86 (5.19)	9.63 (4.50)	10.45 (5.43)	9.92 (4.83)	10.24 (5.19)

Table 6.5 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
T_{CADF}	$ICC = .50$	9.34 (4.41)	9.56 (4.52)	9.22 (4.20)	9.44 (4.36)	9.27 (4.23)	9.49 (4.31)
	$ICC = .26$	9.16 (4.28)	9.36 (4.43)	9.18 (4.14)	9.39 (4.25)	9.30 (4.26)	9.49 (4.38)
	$ICC = .10$	8.75 (3.89)	9.58 (4.83)	9.14 (4.08)	9.37 (4.23)	9.32 (4.26)	9.57 (4.38)
	$ICC = .05$	7.60 (3.13)	10.19 (4.58)	9.10 (4.03)	9.80 (4.75)	9.35 (4.28)	9.62 (4.56)
T_{RADF}	$ICC = .50$	9.96 (5.04)	10.21 (5.18)	9.81 (4.73)	10.06 (4.94)	9.86 (4.77)	10.12 (4.87)
	$ICC = .26$	9.76 (4.87)	9.99 (5.07)	9.76 (4.65)	10.01 (4.79)	9.89 (4.83)	10.11 (4.98)
	$ICC = .10$	9.33 (4.44)	10.37 (6.03)	9.73 (4.60)	10.00 (4.81)	9.93 (4.83)	10.21 (4.99)
	$ICC = .05$	8.21 (3.69)	11.20 (5.64)	9.70 (4.55)	10.55 (5.53)	9.97 (4.85)	10.29 (5.22)
T_{CRADF}	$ICC = .50$	9.37 (4.45)	9.60 (4.56)	9.25 (4.22)	9.47 (4.39)	9.30 (4.25)	9.52 (4.33)
	$ICC = .26$	9.20 (4.32)	9.40 (4.47)	9.21 (4.16)	9.43 (4.27)	9.32 (4.28)	9.51 (4.40)
	$ICC = .10$	8.83 (3.97)	9.70 (5.09)	9.18 (4.11)	9.42 (4.28)	9.35 (4.28)	9.61 (4.40)
	$ICC = .05$	7.82 (3.35)	10.46 (4.93)	9.15 (4.07)	9.89 (4.82)	9.39 (4.30)	9.67 (4.59)
		(b) $J = 100$					
T_{FISHER}	$ICC = .50$	8.29 (4.13)		7.42 (3.55)		8.76 (4.49)	
	$ICC = .26$	8.19 (4.14)		8.47 (4.02)		8.73 (4.47)	
	$ICC = .10$	7.44 (3.64)		8.40 (3.99)		8.62 (4.36)	
	$ICC = .05$	5.76 (2.92)		8.09 (3.74)		8.44 (4.18)	
T_{ML}	$ICC = .50$	12.51 (6.22)		10.26 (4.81)		10.20 (5.09)	

Table 6.5 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
	$ICC = .26$	19.54 (11.10)		12.01 (5.73)		11.21 (5.62)	
	$ICC = .10$	80.96 (61.29)		21.14 (11.75)		15.92 (8.34)	
	$ICC = .05$	192.24 (96.31)		47.53 (41.23)		26.96 (19.25)	
T_{RML}	$ICC = .50$	9.59 (4.76)	9.69 (4.80)	9.41 (4.38)	9.51 (4.42)	9.71 (4.85)	9.81 (4.90)
	$ICC = .26$	9.91 (5.48)	10.06 (5.64)	9.48 (4.49)	9.59 (4.54)	9.74 (4.90)	9.83 (4.94)
	$ICC = .10$	11.58 (8.01)	13.98 (10.43)		10.05 (5.50)	9.91 (5.20)	10.02 (5.27)
	$ICC = .05$	8.99 (5.74)	14.32 (8.13)	11.33 (9.47)	12.35 (11.12)	10.63 (7.24)	10.93 (7.85)
T_{ADF}	$ICC = .50$	10.71 (5.44)	11.02 (5.68)	10.42 (5.16)	10.66 (5.25)	10.73 (5.57)	11.00 (5.86)
	$ICC = .26$	10.49 (5.29)	10.81 (5.52)	10.39 (5.09)	10.66 (5.21)	10.71 (5.64)	10.87 (5.62)
	$ICC = .10$	9.23 (4.15)	11.54 (6.03)	10.25 (4.85)	10.68 (5.30)	10.63 (5.59)	10.88 (5.71)
	$ICC = .05$	6.86 (2.91)	10.41 (4.39)	9.96 (4.55)	11.60 (6.57)	10.58 (5.48)	11.47 (6.53)
T_{CADF}	$ICC = .50$	9.46 (4.23)	9.69 (4.38)	9.24 (4.08)	9.43 (4.14)	9.46 (4.35)	9.65 (4.53)
	$ICC = .26$	9.28 (4.15)	9.53 (4.30)	9.22 (4.03)	9.43 (4.11)	9.44 (4.38)	9.57 (4.37)
	$ICC = .10$	8.32 (3.38)	10.08 (4.62)	9.12 (3.87)	9.44 (4.17)	9.38 (4.33)	9.57 (4.40)
	$ICC = .05$	6.34 (2.49)	9.27 (3.50)	8.89 (3.67)	10.10 (4.87)	9.35 (4.26)	9.99 (4.86)
T_{RADF}	$ICC = .50$	10.85 (5.60)	11.20 (5.89)	10.56 (5.29)	10.82 (5.39)	10.89 (5.74)	11.17 (6.06)
	$ICC = .26$	10.69 (5.50)	11.04 (5.76)	10.55 (5.24)	10.84 (5.38)	10.88 (5.84)	11.09 (5.88)
	$ICC = .10$	9.64 (4.61)	12.05 (6.59)	10.45 (5.04)	10.93 (5.58)	10.84 (5.84)	11.13 (6.04)

Table 6.5 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
	$ICC = .05$	7.44 (3.50)	11.23 (5.33)	10.26 (4.80)	12.11 (7.05)	10.87 (5.80)	11.86 (6.94)
T_{CRADF}	$ICC = .50$	9.55 (4.31)	9.81 (4.49)	9.33 (4.15)	9.53 (4.22)	9.56 (4.45)	9.77 (4.63)
	$ICC = .26$	9.43 (4.27)	9.69 (4.44)	9.32 (4.12)	9.55 (4.21)	9.55 (4.49)	9.72 (4.52)
	$ICC = .10$	8.61 (3.69)	10.43 (4.97)	9.26 (3.99)	9.61 (4.33)	9.52 (4.46)	9.74 (4.58)
	$ICC = .05$	6.82 (2.94)	9.87 (4.13)	9.12 (3.83)	10.45 (5.13)	9.55 (4.46)	10.26 (5.10)
				(b) $J = 50$			
T_{FISHER}	$ICC = .50$	7.58 (3.69)		7.61 (3.49)		7.91 (4.15)	
	$ICC = .26$	7.28 (3.51)		7.61 (3.51)		7.85 (4.11)	
	$ICC = .10$	5.95 (3.01)		7.44 (3.58)		7.71 (4.02)	
	$ICC = .05$	4.21 (2.43)		6.78 (3.43)		7.38 (3.79)	
T_{ML}	$ICC = .50$	13.53 (8.11)		10.64 (4.97)		10.42 (5.19)	
	$ICC = .26$	24.65 (20.46)		12.85 (6.41)		11.63 (5.87)	
	$ICC = .10$	93.47 (53.61)		26.78 (19.64)		17.72 (9.96)	
	$ICC = .05$	161.61 (67.48)		64.32 (46.88)		35.03 (28.11)	
T_{RML}	$ICC = .50$	10.29 (6.01)	10.52 (6.15)	9.81 (4.45)	10.01 (4.54)	10.03 (5.06)	10.22 (5.14)
	$ICC = .26$	11.83 (9.52)	12.49 (10.37)		10.27 (4.95)	10.16 (5.20)	10.38 (5.33)
	$ICC = .10$	12.21 (7.50)	16.96 (9.84)	11.88 (8.95)	12.56 (9.77)	10.77 (6.06)	11.03 (6.23)
	$ICC = .05$	8.99 (5.74)	13.12 (8.02)	11.33 (9.47)	16.95 (12.64)	10.63 (7.24)	14.32 (8.13)

Table 6.5 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
T_{ADF}	$ICC = .50$	11.70 (5.89)	11.92 (6.12)	11.55 (5.46)	11.81 (5.55)	11.81 (5.78)	12.13 (6.02)
	$ICC = .26$	11.43 (5.52)	12.32 (6.57)	11.62 (5.59)	11.97 (5.83)	11.90 (5.90)	12.27 (6.21)
	$ICC = .10$	9.01 (3.83)	12.12 (5.16)	11.49 (5.64)	12.43 (6.51)	11.93 (6.11)	12.24 (6.28)
	$ICC = .05$	5.75 (2.56)	8.93 (3.54)	10.33 (4.60)	12.76 (6.01)	11.66 (5.82)	10.41 (4.39)
T_{CADF}	$ICC = .50$	9.11 (3.52)	9.23 (3.56)	9.04 (3.40)	9.20 (3.44)	9.18 (3.54)	9.36 (3.63)
	$ICC = .26$	8.97 (3.33)	9.46 (3.69)	9.07 (3.44)	9.29 (3.52)	9.23 (3.58)	9.44 (3.68)
	$ICC = .10$	7.44 (2.62)	9.44 (3.20)	8.99 (3.47)	9.51 (3.83)	9.23 (3.66)	9.41 (3.72)
	$ICC = .05$	5.05 (2.01)	7.40 (2.47)	8.29 (3.04)	9.77 (3.64)	9.08 (3.56)	9.27 (3.50)
T_{RADF}	$ICC = .50$	12.70 (7.27)	12.99 (7.55)	12.33 (6.24)	12.58 (6.39)	12.72 (6.98)	13.05 (7.14)
	$ICC = .26$	12.50 (6.40)	13.49 (7.54)	12.49 (6.56)	12.88 (6.75)	12.85 (7.16)	13.22 (7.41)
	$ICC = .10$	9.88 (4.76)	13.09 (6.20)	12.55 (6.61)	13.66 (7.71)	12.92 (7.17)	13.26 (7.34)
	$ICC = .05$	6.58 (3.37)	10.09 (4.66)	11.46 (5.70)	14.10 (7.25)	12.70 (6.86)	11.23 (5.33)
T_{CRADF}	$ICC = .50$	9.58 (3.94)	9.74 (4.00)	9.43 (3.70)	9.58 (3.74)	9.60 (3.97)	9.79 (4.04)
	$ICC = .26$	9.53 (3.69)	10.04 (4.04)	9.51 (3.79)	9.73 (3.85)	9.67 (4.02)	9.88 (4.09)
	$ICC = .10$	7.94 (3.08)	9.92 (3.63)	9.53 (3.84)	10.11 (4.25)	9.71 (4.03)	9.91 (4.08)
	$ICC = .05$	5.63 (2.49)	8.10 (3.02)	8.91 (3.54)	10.42 (4.12)	9.61 (3.92)	9.87 (4.13)

Tables 6.5 and 6.6 display test statistic means and standard deviations for all six test statistics, as well as T_{FISHER}, the ADF test statistic that uses Γ_{FISHER}. Under normality, Γ_{FISHER} is correctly specified, and so T_{FISHER} will have the correct distribution as an asymptotic central chi square variate. The five test statistics that use either $\hat{\Gamma}_{GEE}$ or $\hat{\Gamma}_{BOOT}$, namely T_{RML}, T_{ADF}, T_{CADF}, T_{RADF} and T_{CRADF}, are presented side by side for direct comparison of the GEE and bootstrap approaches.

Turning first to the small models (Table 6.5), it is clear that T_{ML} only behaves like a central chi-square variate on 9 degrees of freedom when the within-group sample sizes are large and the ICCs are high. This can be seen, for example, in Table 6.5a, when $n = 50$ and $ICC = .50$, T_{ML} has a mean of 9.76 and a standard deviation of 4.52. There is also some evidence that, for smaller within-group sample sizes or lower ICCs, T_{ML} does not achieve the correct chi-square distribution, even asymptotically. For example, in Table 6.5, for $ICC = .10$ and $n = 10$, as the level 2 sample sizes increase from 50 to 200, the mean of T_{ML} changes from 93.47 to 63.00, but is nowhere near the theoretical mean of 9.

T_{RML} does correct the T_{ML} test statistic, and the means T_{RML} are systematically closer to the theoretical mean of 9 than the means of T_{ML}. This is true for the GEE and bootstrap-based versions of this statistic. However, particularly with small samples, the means of T_{RML} are still higher than anticipated (Table 6.5c).

There is evidence that T_{RML}, T_{ADF}, T_{CADF}, T_{RADF} and T_{CRADF} do converge to the correct distribution across all ICC and within-group sample size conditions as the number of groups increases. This is true for both the GEE and bootstrap based test statistics. For each of these statistics, the empirical means get closer to the theoretical mean (9) as the level 2 sample sizes increase from 50 to 200. For example, at $ICC = .50$ and $n = 10$, as J goes from 50 to 200, the means of T_{RADF} go from 12.70 to 9.96 for the GEE based test statistics, and from 12.99 to 10.21 for the bootstrap based test statistics. For small level 2 sample sizes (6.5c),

T_{CRADF} and T_{CADF} have smaller variance than expected (as reported in Bentler and Yuan (1999) for the conventional version of these statistics).

As J, the level-2 sample size, increases, the performance of the bootstrap-based test statistics becomes very similar to the GEE-based test statistics. This can be seen in Table 6.5 because the differences between the GEE and bootstrap based estimates for a given ICC/within group sample size combination decrease as J goes from 50 to 200. However, the GEE-based test statistics show better performance at smaller sample sizes and at lower ICC conditions. For example, in Table 6.5a, with $J = 50$, the means of the bootstrap based test statistics T_{RADF} and T_{ADF} are systematically larger than those of the GEE based test statistics. The magnitude of this difference increases as the ICC decreases: at $ICC = .50$ and $n = 10$, the difference averages around .2. At $ICC = .10$, and $n = 10$, the difference averages around 2. On the other hand, at $ICC = .50$ and $n = 50$, the difference averages around .2 for all ICCs.

Turning to the large models (Table 6.6), many of the same patterns are visible. Specifically, in nearly every condition, the mean of T_{ML} is too high. With $J = 100$, $ICC = .05$ and $n = 10$ (Table 6.6c), the mean of T_{ML} is over 1,000, even though the theoretical mean of the χ^2_{54} distribution is 54. T_{RML} corrects the T_{ML} test statistic, and the means T_{RML} are systematically closer to the theoretical mean of 54 than the means of T_{ML}. This is true for the GEE and bootstrap-based versions of this statistic. However, particularly with small samples, T_{RML} is unable to scale the T_{ML} test statistic adequately (for example, in Table 6.6c).

The test statistics T_{RML}, T_{ADF}, T_{CADF}, T_{RADF} and T_{CRADF} show a pattern of convergence as J goes from 50 to 200, in that the means of these statistics steadily approach 54. However, even with 200 groups, the means and variances of T_{ADF} and T_{CADF} are too large in all ICC and within group sample size conditions (Table 6.6a). This suggests that when the model is sufficiently large, the number of groups would have to be enormous in order for T_{ADF} and T_{CADF} to provide correct

inferences. This is consistent with previously reported results (e.g., Curran et al., 1996; Hu et al., 1992; B. O. Muthén & Kaplan, 1985, 1992; Yuan & Bentler, 2003; Bentler & Yuan, 1999; Powell & Schafer, 2001). For small level 2 sample sizes (Table 6.5b and 6.5c), T_{CRADF} and T_{CADF} have smaller variance than expected.

Table 6.6: Test statistic means and standard deviations, $df = 54$.

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
				(a) $J = 200$			
T_{FISHER}	$ICC = .50$	51.90 (10.83)		52.59 (10.74)		52.36 (10.91)	
	$ICC = .26$	51.60 (10.80)		52.47 (10.86)		52.43 (10.90)	
	$ICC = .10$	49.90 (10.31)		52.03 (11.13)		52.63 (10.91)	
	$ICC = .05$	43.51 (8.85)		51.10 (10.91)		52.47 (10.79)	
T_{ML}	$ICC = .50$	73.45 (14.97)		61.17 (12.23)		58.67 (11.49)	
	$ICC = .26$	113.19 (32.07)		70.75 (14.56)		64.35 (12.70)	
	$ICC = .10$	445.88 (163.04)		121.20 (30.51)		91.14 (19.63)	
	$ICC = .05$	1233.87 (299.86)		262.93 (109.38)		150.01 (50.77)	
T_{RML}	$ICC = .50$	56.46 (11.50)	56.79 (11.56)	56.11 (11.11)	56.36 (11.18)	55.79 (10.87)	56.11 (10.97)
	$ICC = .26$	59.55 (16.50)	60.14 (16.89)	56.48 (11.54)	56.78 (11.67)	56.19 (11.05)	56.43 (11.11)
	$ICC = .10$	77.47 (28.13)	88.16 (33.80)	59.31 (14.65)	59.84 (14.91)	58.05 (12.39)	58.39 (12.49)
	$ICC = .05$	76.74 (21.20)	110.06 (28.77)	70.37 (28.15)	74.79 (31.55)	62.95 (20.48)	63.94 (21.48)
T_{ADF}	$ICC = .50$	80.03 (18.68)	89.37 (21.54)	80.02 (18.92)	89.33 (21.64)	80.12 (18.29)	89.47 (21.69)
	$ICC = .26$	80.01 (19.00)	90.37 (23.09)	79.65 (18.51)	88.78 (21.20)	80.38 (18.48)	89.40 (21.66)
	$ICC = .10$	75.26 (16.39)	98.80 (25.13)	79.06 (17.89)	89.09 (21.63)	80.53 (18.70)	89.47 (20.89)
	$ICC = .05$	61.24 (11.19)	9.70 (18.16)	78.12 (17.85)	94.65 (26.32)	79.97 (18.35)	91.92 (24.95)

Table 6.6 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
T_{CADF}	$ICC = .50$	56.45 (9.38)	60.92 (10.13)	56.42 (9.56)	60.89 (10.27)	56.53 (9.05)	60.97 (9.99)
	$ICC = .26$	56.42 (9.51)	61.31 (10.58)	56.26 (9.37)	60.65 (10.06)	56.65 (9.13)	60.94 (9.99)
	$ICC = .10$	54.10 (8.44)	65.11 (10.83)	55.99 (9.12)	60.78 (10.15)	56.72 (9.24)	61.02 (9.70)
	$ICC = .05$	46.55 (6.49)	62.72 (8.34)	55.51 (9.06)	63.11 (11.62)	56.45 (9.16)	61.96 (10.82)
T_{RADF}	$ICC = .50$	81.20 (19.29)	91.12 (22.66)	81.26 (19.61)	91.23 (22.73)	81.19 (18.74)	91.13 (22.37)
	$ICC = .26$	81.75 (20.09)	93.21 (24.99)	81.02 (19.26)	90.89 (22.52)	81.55 (18.98)	91.07 (22.49)
	$ICC = .10$	78.40 (18.18)	105.26 (29.35)	80.73 (18.72)	91.65 (22.95)	81.98 (19.51)	91.52 (22.17)
	$ICC = .05$	65.96 (13.13)	103.85 (22.06)	80.43 (19.01)	98.62 (28.69)	81.86 (19.36)	94.77 (26.47)
T_{CRADF}	$ICC = .50$	56.92 (9.55)	61.59 (10.42)	56.93 (9.77)	61.62 (10.59)	56.97 (9.18)	61.62 (10.16)
	$ICC = .26$	57.16 (9.85)	62.43 (11.11)	56.83 (9.61)	61.49 (10.44)	57.13 (9.27)	61.59 (10.21)
	$ICC = .10$	55.57 (9.07)	67.58 (11.95)	56.72 (9.39)	61.82 (10.53)	57.32 (9.47)	61.82 (10.05)
	$ICC = .05$	49.11 (7.32)	67.44 (9.39)	56.55 (9.44)	64.66 (12.22)	57.26 (9.46)	63.10 (11.20)
		(b) $J = 100$					
T_{FISHER}	$ICC = .50$	50.09 (10.34)		52.04 (10.97)		51.14 (10.78)	
	$ICC = .26$	49.14 (9.97)		51.84 (11.07)		51.05 (10.74)	
	$ICC = .10$	44.55 (8.97)		51.08 (10.82)		50.69 (10.55)	
	$ICC = .05$	34.67 (7.62)		48.62 (9.92)		49.67 (10.08)	
T_{ML}	$ICC = .50$	78.61 (16.35)		64.66 (12.75)		61.08 (11.66)	

Table 6.6 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
	$ICC = .26$	136.80 (54.22)		76.32 (16.39)		67.56 (13.37)	
	$ICC = .10$	553.57 (153.68)		151.83 (63.90)		100.85 (29.09)	
	$ICC = .05$	1144.91 (205.66)		341.58 (110.40)		190.39 (78.90)	
T_{RML}	$ICC = .50$	60.24 (12.45)	60.83 (12.55)	59.41 (11.57)	59.96 (11.69)	58.42 (11.22)	59.00 (11.29)
	$ICC = .26$	70.60 (27.18)	73.06 (28.90)	60.76 (12.92)	61.37 (13.00)	59.17 (11.77)	59.77 (11.88)
	$ICC = .10$	92.50 (26.73)	115.53 (33.23)	72.71 (29.84)	75.00 (31.45)	63.67 (18.20)	64.48 (18.82)
	$ICC = .05$	73.08 (20.97)	117.93 (27.85)	89.19 (29.85)	100.69 (34.23)	77.78 (31.76)	81.95 (34.40)
T_{ADF}	$ICC = .50$	126.57 (33.90)	135.75 (35.96)	127.85 (30.91)	137.78 (33.10)	126.63 (32.33)	136.19 (34.77)
	$ICC = .26$	123.71 (33.93)	139.02 (39.86)	126.36 (30.77)	136.20 (32.92)	126.14 (32.21)	136.41 (34.59)
	$ICC = .10$	103.37 (24.69)	130.26 (30.09)	123.19 (31.31)	138.05 (37.34)	125.46 (31.60)	137.58 (36.72)
	$ICC = .05$	71.86 (16.87)	95.61 (19.10)	113.90 (28.08)	135.08 (33.08)	122.52 (30.54)	140.67 (38.25)
T_{CADF}	$ICC = .50$	54.60 (6.40)	56.31 (6.27)	55.01 (5.81)	56.80 (5.85)	54.67 (6.28)	56.43 (6.16)
	$ICC = .26$	54.02 (6.47)	56.71 (6.78)	54.72 (5.85)	56.53 (5.78)	54.59 (6.23)	56.48 (6.11)
	$ICC = .10$	49.87 (5.78)	55.53 (5.57)	54.03 (6.14)	56.67 (6.38)	54.50 (6.04)	56.63 (6.24)
	$ICC = .05$	41.11 (5.40)	48.17 (4.79)	52.20 (5.87)	56.32 (5.80)	53.94 (6.01)	57.12 (6.32)
T_{RADF}	$ICC = .50$	143.50 (45.62)	154.53 (48.58)	144.83 (40.46)	156.19 (43.60)	142.42 (42.70)	152.25 (44.79)
	$ICC = .26$	144.58 (46.83)	161.54 (52.54)	144.95 (40.85)	156.56 (44.07)	142.84 (42.32)	153.94 (44.59)
	$ICC = .10$	123.84 (34.99)	155.72 (42.07)	143.93 (41.81)	161.44 (48.88)	143.83 (41.79)	157.27 (46.64)

Table 6.6 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
	$ICC = .05$	86.06 (23.37)	115.01 (27.61)	137.14 (40.20)	161.64 (44.37)	143.44 (41.20)	164.23 (50.35)
T_{CRADF}	$ICC = .50$	56.91 (7.22)	58.66 (7.03)	57.40 (6.44)	59.14 (6.43)	56.83 (7.04)	58.43 (6.82)
	$ICC = .26$	57.03 (7.33)	59.55 (7.36)	57.40 (6.55)	59.19 (6.41)	56.93 (6.95)	58.72 (6.72)
	$ICC = .10$	53.67 (6.70)	59.13 (6.30)	57.17 (6.73)	59.74 (6.80)	57.15 (6.75)	59.18 (6.77)
	$ICC = .05$	45.03 (6.32)	52.20 (5.57)	56.04 (6.75)	59.97 (6.25)	57.12 (6.66)	60.12 (6.79)
				(b) $J = 50$			
T_{ML}	$ICC = .50$	89.03 (21.13)		69.65 (13.55)		65.74 (12.58)	
	$ICC = .26$	182.84 (64.49)		86.86 (22.63)		74.41 (14.99)	
	$ICC = .10$	584.55 (108.73)		214.54 (72.76)		134.04 (49.21)	
	$ICC = .05$	848.61 (115.48)		433.28 (107.79)		266.90 (77.61)	
T_{RML}	$ICC = .50$	67.59 (15.65)	69.29 (16.27)	64.30 (12.55)	65.57 (12.81)	63.33 (12.09)	64.54 (12.33)
	$ICC = .26$	91.78 (32.45)	99.26 (35.54)	68.76 (17.69)	70.41 (18.16)	65.27 (13.13)	66.52 (13.46)
	$ICC = .10$	96.78 (23.82)	134.45 (29.33)	99.51 (34.03)	108.00 (37.05)	82.93 (30.60)	86.43 (32.13)
	$ICC = .05$	60.35 (20.27)	106.69 (27.20)	109.10 (29.21)	134.94 (35.32)	106.17 (31.95)	118.69 (35.63)
T_{RADF}	$ICC = .50$		534.09 (192.46)		563.25 (195.35)		565.23 (203.48)
	$ICC = .26$		463.85 (147.62)		551.27 (189.19)		559.70 (202.34)
	$ICC = .10$		249.83 (71.69)		436.85 (123.34)		514.24 (175.05)
	$ICC = .05$		141.38 (41.74)		312.77 (90.29)		407.57 (125.86)

Table 6.6 – continued from previous page

		Group Sizes					
		10		30		50	
		GEE	Bootstrap	GEE	Bootstrap	GEE	Bootstrap
T_{CRADF}	$ICC = .50$		43.64 (1.37)		43.87 (1.27)		43.85 (1.35)
	$ICC = .26$		43.13 (1.41)		43.81 (1.24)		43.80 (1.37)
	$ICC = .10$		39.83 (1.97)		42.95 (1.27)		43.49 (1.48)
	$ICC = .05$		35.26 (2.78)		41.27 (1.49)		42.57 (1.50)

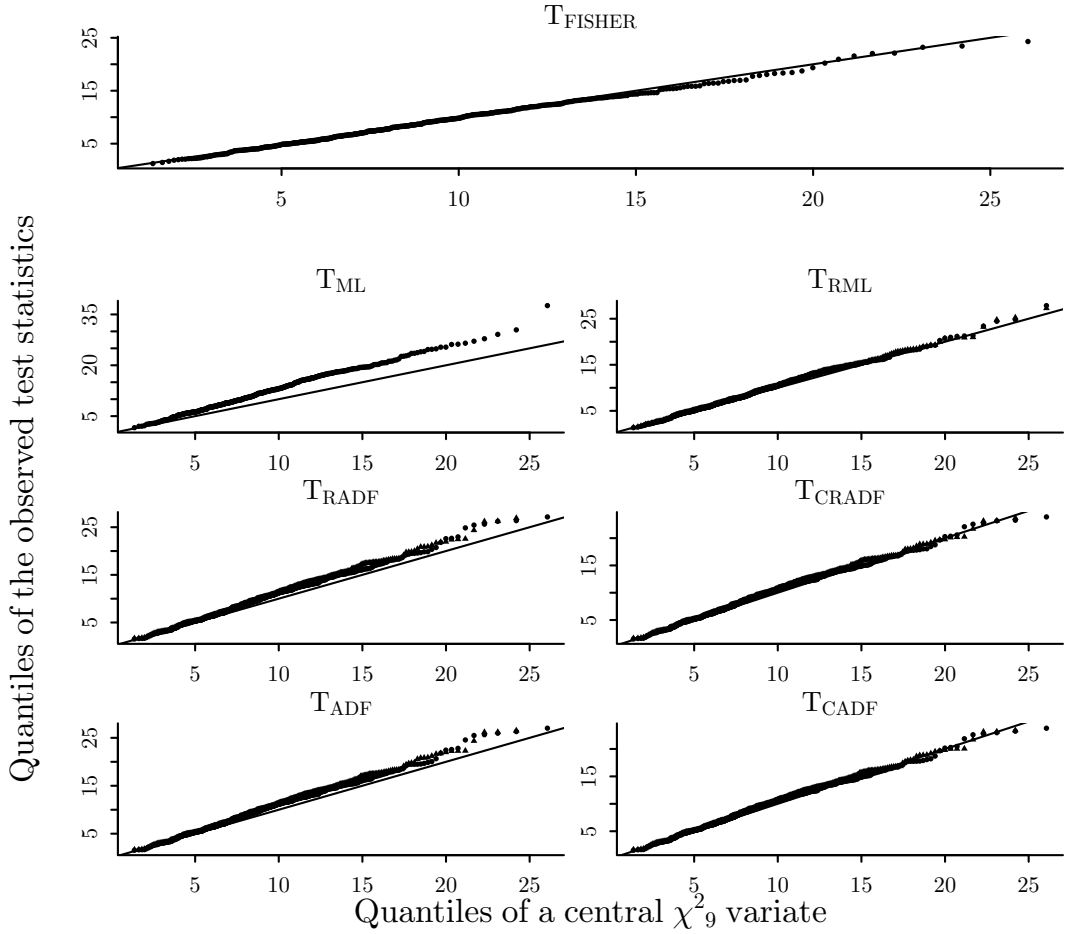
There is a clearer separation in the performance of the GEE and bootstrap-based test statistics with the larger model, particularly as regards the corrected test statistics. For example, as can be seen in Table 6.6a, the means of the bootstrap-based T_{RADF} and T_{ADF} test statistics are systematically larger than their GEE-based counterparts. As was the case with the small models ($df = 9$), the magnitude of this difference is related to ICC: as ICC decreases the difference increases. Because the bootstrap-based test statistics T_{RADF} and T_{ADF} are systematically larger than their GEE-based counterparts, at smaller sample sizes, the bootstrap-based T_{CRADF} is unable to appropriately correct the performance of T_{RADF}.

When the sample size is small ($J = 50$) and the model is large ($df = 54$), the GEE based test statistics T_{RADF} and T_{CRADF} are not estimable because $[\dot{\sigma}_c(\hat{\theta})^T \hat{\Gamma}_{GEE} \dot{\sigma}_c(\hat{\theta})]^{-1}$ (Equation 3.19) is not invertible. The bootstrap-based versions of these test statistics are computable. However, the performance of these statistics is quite poor. The bootstrap-based T_{RADF} has a mean far higher than the theoretical value of 54 (6.6c). On the other hand, the corrected test statistic T_{CRADF} has a mean that is too low. In fact, for small sample sizes and large models, as presented in Table 6.6c, no test statistics show good performance, and none of the statistics have empirical means that reflect the theoretical chi square distribution.

6.2.4.3 Q-Q plots for test statistic distributions

While the Type I error rates, means, and standard deviations are important considerations, Q-Q plots provide a way to assess the overall distribution of the test statistic, particularly the performance in the tails of the distribution. Figures 6.13 and 6.14 present Q-Q plots for two different simulation conditions with small models ($df = 9$). Figure 6.13 shows a Q-Q plot for a condition with $ICC = .26$, $n = 30$ and $J = 200$. Figure 6.14 shows a Q-Q plot for a condition with $ICC = .26$,

Figure 6.13: Q-Q plot for $df = 9$, $ICC = .26$, $J = 200$, $n = 30$

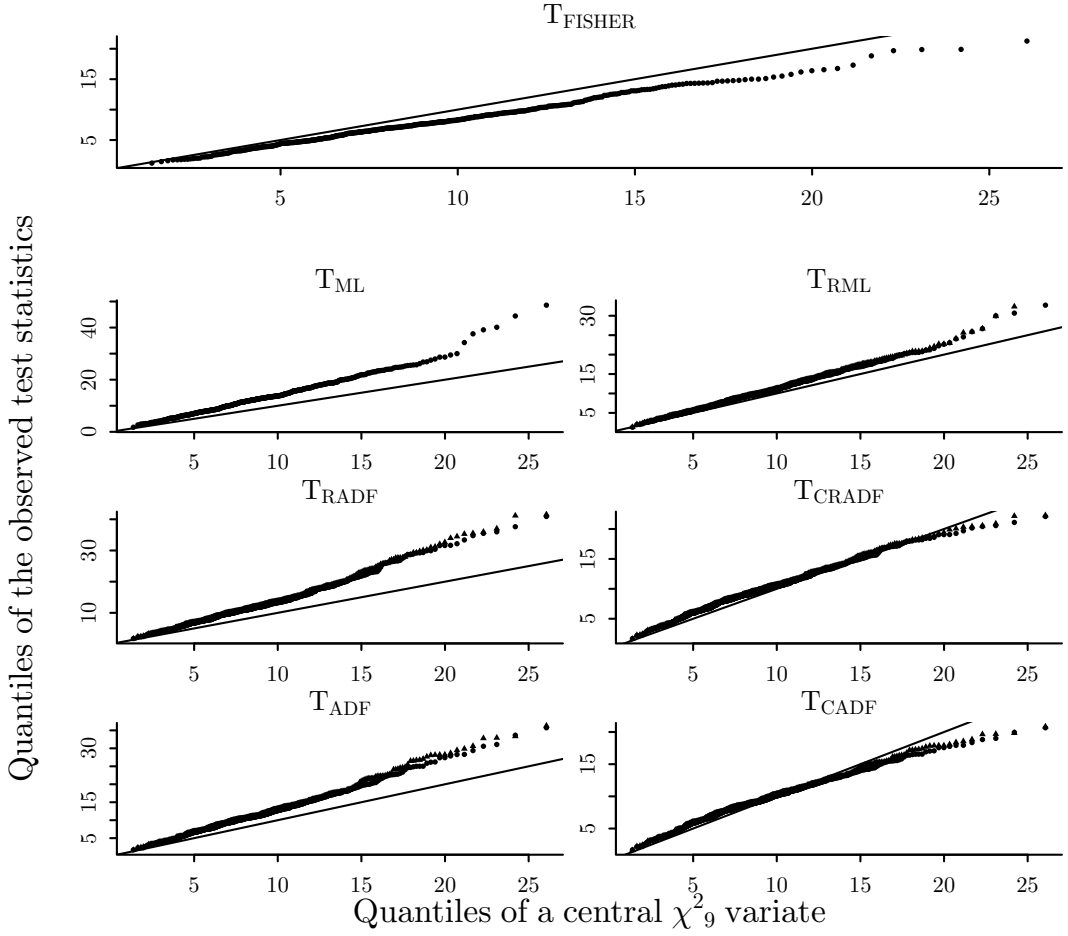


Legend: ▲ =Bootstrap, • =GEE

$n = 30$ and $J = 50$. The solid line in each plot represents the theoretical chi-square distribution. In other words, the empirical distribution more closely follows the theorized chi-square distribution if the points follow the line closely. The GEE-based estimates are represented by black circles (•), and the cluster bootstrap based estimates are represented by black triangles (▲). For reference, a Q-Q plot for T_{FISHER} is included when available. Complete Q-Q plots, for all 72 simulation conditions, are available in Appendix A.

The Q-Q plots reinforce many of the results presented in Tables 6.3-6.6 above.

Figure 6.14: Q-Q plot $df = 9$, $ICC = .26$, $J = 50$, $n = 30$



Legend: ▲ =Bootstrap, ● =GEE

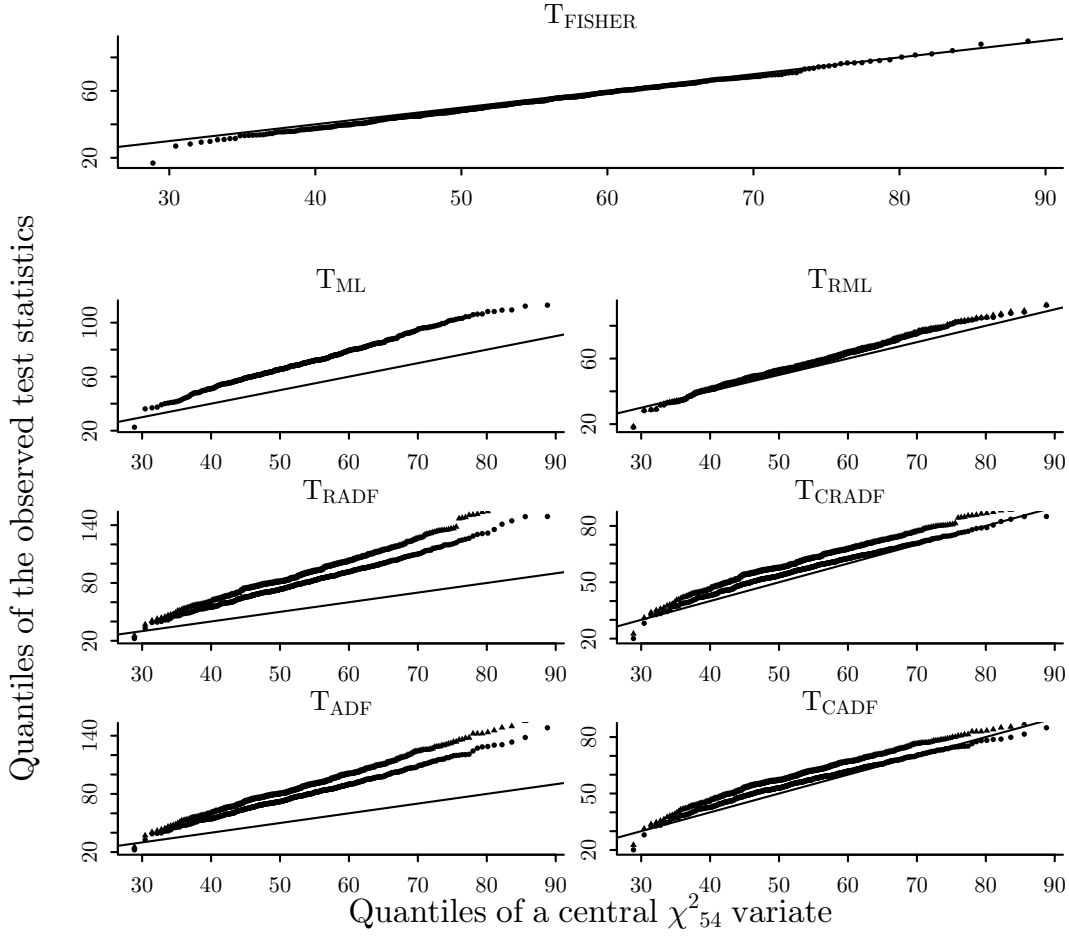
T_{ML} does not show the appropriate distribution under either set of conditions. In Figure 6.14, when the level-2 sample size is smaller, observed T_{ML} values are systematically greater than values expected based on the theoretical χ^2_9 distribution. This is particularly true for the upper tail of the distribution. The GEE and bootstrap-based versions of T_{RML} perform better than T_{ML}, in that the observed test statistics are closer to the solid line representing the theoretical chi-square distribution. With small level-2 sample sizes T_{RADF} and T_{ADF} deviate from the expected distribution, as do T_{CRADF} and T_{CADF}. This is particularly true in the upper tail.

However, as sample sizes increase, the empirical distributions of T_{RML}, T_{ADF}, T_{RADF}, T_{CADF} and T_{CRADF} more closely match the theoretical distributions. In Figure 6.13, these test statistics are well-behaved, in that their empirical distributions closely match the theoretical distributions. This suggests more evidence that these statistics are correctly distributed asymptotically. In fact, the same pattern can be observed for T_{FISHER}, which is known to be asymptotically correct. At $J = 50$, (Figure 6.14), the observed values of T_{FISHER} are systematically too low. However, at $J = 200$, the observed values follow the theoretical distribution closely.

Based on Figures 6.13 and 6.14, it would appear that the overall distribution of the bootstrap test statistics is very similar to the overall distribution of the GEE test statistics. The plots of the bootstrap test statistics T_{RML}, T_{ADF}, T_{RADF}, T_{CADF} and T_{CRADF}—represented by the ▲ symbols—and the GEE test statistics—represented by the • symbols—are nearly indistinguishable from one another at both $J = 50$ and $J = 200$. This provides some additional evidence that the bootstrap test statistics converge to the appropriate distribution as sample size increases.

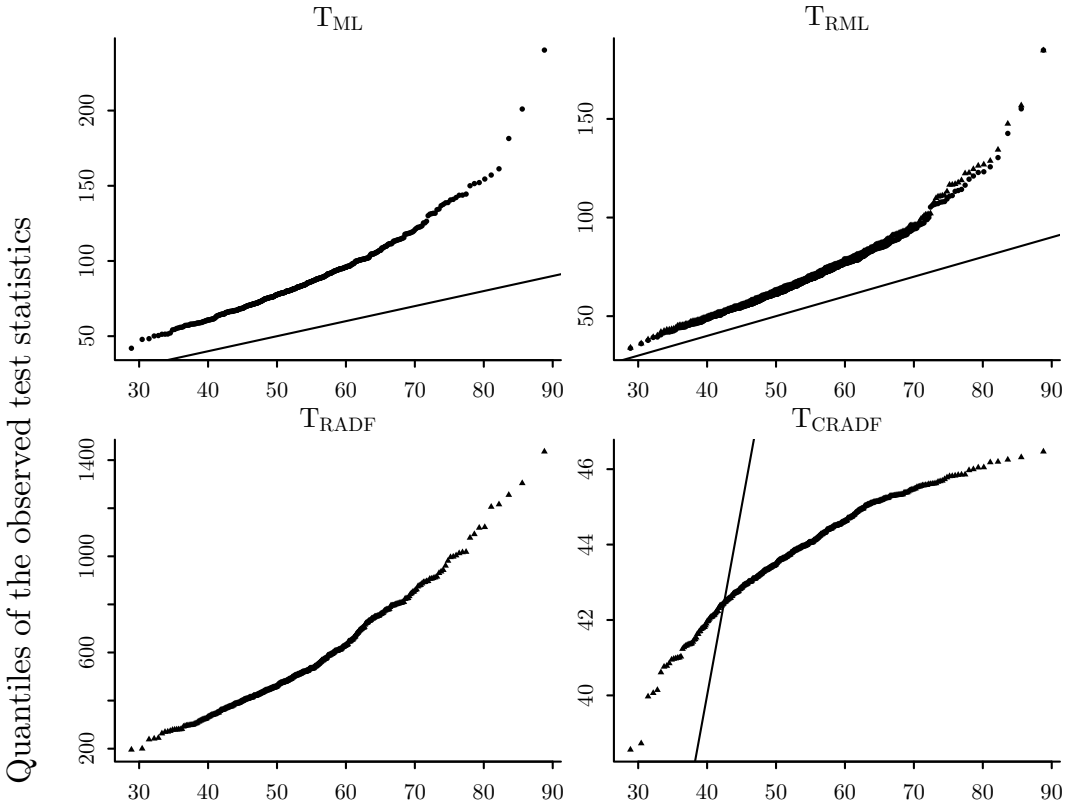
Turning to the large models ($df = 54$), Figure 6.15 and 6.16 present Q-Q plots for two different simulation conditions with small models ($df = 54$). Figure 6.13

Figure 6.15: Q-Q plot for $df = 54$, $ICC = .26$, $J = 200$, $n = 30$



Legend: ▲ =Bootstrap, ● =GEE

Figure 6.16: Q-Q plot $df = 54$, $ICC = .26$, $J = 50$, $n = 30$



shows a Q-Q plot for a condition with $ICC = .26$, $n = 30$ and $J = 200$. Figure 6.14 shows a Q-Q plot for a condition with $ICC = .lor26$, $n = 30$ and $J = 50$. Here, T_{ML}, T_{RADF} and T_{ADF} all perform terribly, and are not close to the appropriate distribution, even with sample sizes of 200 (Figure 6.15). Thus, the larger model conditions make it clear that, in order for the test statistics to behave as a central chi square variate, sufficiently large level 2 sample sizes are needed.

The difference in performance between the GEE and bootstrap methods are more noticeable with the large model (Figure 6.15) . Whereas with the small models, the GEE and bootstrap plots essentially overlapped, with the large models, there is a detectible difference between the two plots, and the bootstrap-based estimates are consistently higher than the GEE-based estimates.

When the sample size is small ($J = 50$) and the model is large ($df = 54$), none of the estimable statistics performs well. T_{RADF} is far too large, and the distribution of T_{CRADF} does not match the theoretical distribution, and deviates greatly from the correct behavior.

6.2.5 Estimation of Γ_B

The results presented above suggest that for sufficiently large sample sizes—particularly relative to the size of the model—the GEE and bootstrap approaches perform very similarly in terms of parameter bias, parameter mean square error and test statistic performance. The test statistic means, standard deviations and rejection rates for the bootstrap approach are similar to those based on the GEE approach for sufficiently large sample sizes, and both GEE and boostrap-based test statistics show evidence of converging to the proper chi-square distributions. However, for moderate sample sizes, even when T_{FISHER} is appropriately distributed, the bootstrap-based test statics are systematically larger than the GEE-based test statistics.

One hypothesis as to why this occurs is that Γ is better estimated under the GEE

method than the bootstrap for small sample sizes. In the context of conventional factor analysis, Hu et al. (1992) suggested that he poor performance of the ADF method was related to the poor estimation of the elements of Γ, and empirical results in Yung and Bentler (1994) supported this claim.

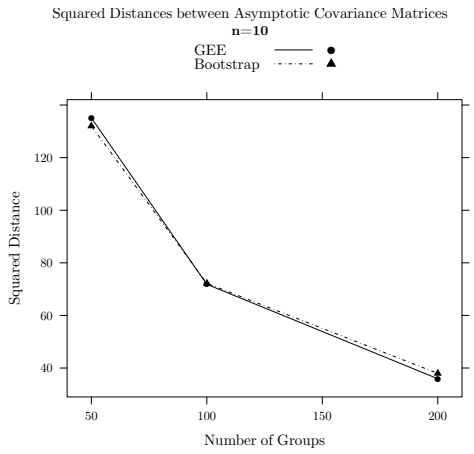
Since Γ_{FISHER} is correct under the assumption of multivariate normality, the consistency and accuracy of $\hat{\Gamma}_{GEE}$ and $\hat{\Gamma}_{BOOT}$ were assessed through their squared distances from Γ_{FISHER} , as was done in previous research Yuan and Hayashi (2006); Yung and Bentler (1994) in conventional confirmatory factor analysis. If these estimated covariance matrices are consistent, D^2 should approach zero. Figures 6.5- 6.12 display results for all model conditions.

The plots in Figures 6.17- 6.24 suggest that as the number groups increases, the squared distances between the asymptotic covariance matrices decreases. The fact that the differences disappear asymptotically is reflective of the consistency property (Yuan & Hayashi, 2006, p. 16). That is, both the GEE and bootstrap based estimators result in consistent estimates of the asymptotic covariance matrix. Several other things are worth noting from these plots. As the ICC decreases, the values of D^2 increase. For example, at $ICC = .50$ (Table 6.17), the squared distances are in the order of 100. At $ICC = .05$ (Table 6.20), the squared distances are range from around 300 to around 5,000. As within group sample sizes decrease, a similar pattern can be found. This occurs for both model sizes. This indicates that the ICC and the within group sample size play a role in the quality of the estimation of the asymptotic covariance matrix, particularly at small sample sizes and low ICCs.

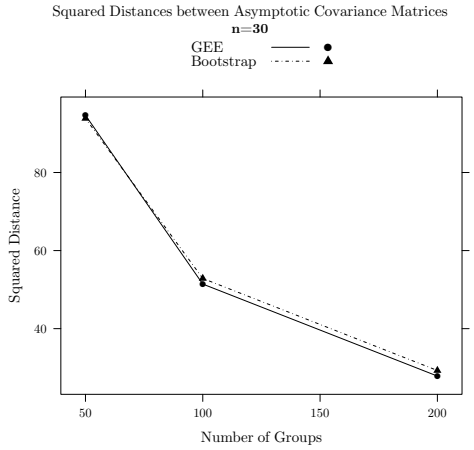
The GEE-based estimator consistently has smaller squared distances for 100 and 200 groups, regardless of ICC or group size conditions, and this may help to explain the better performance of the test statistics under these conditions. This pattern does not hold for the $J = 50$ condition, however. In other words, the squared distances are systematically lower for the bootstrap-based approach than they are

Figure 6.17: D^2 plots: asymptotic covariance matrices, $df = 9$ $ICC = .50$

(a)



(b)



(c)

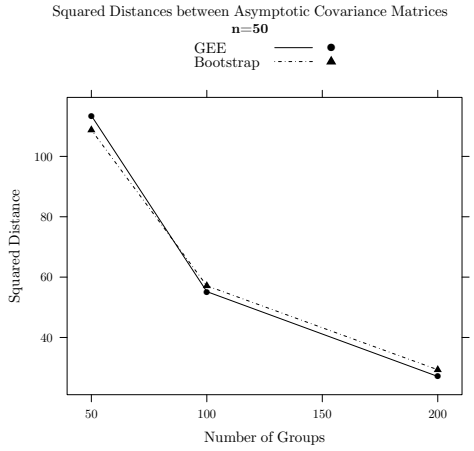
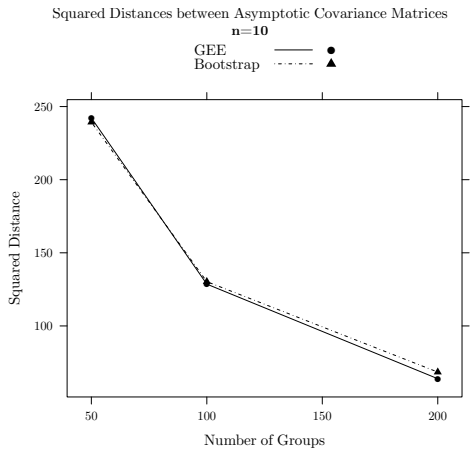
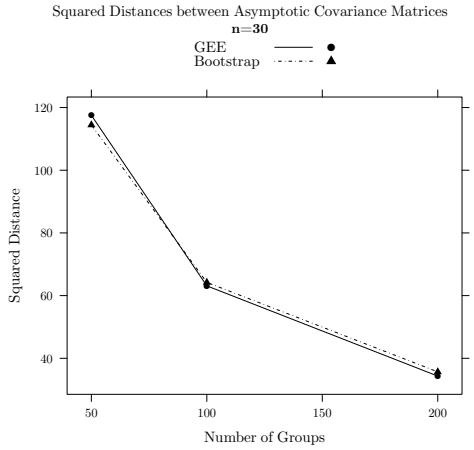


Figure 6.18: D^2 plots: asymptotic covariance matrices, $df = 9$ $ICC = .26$

(a)



(b)



(c)

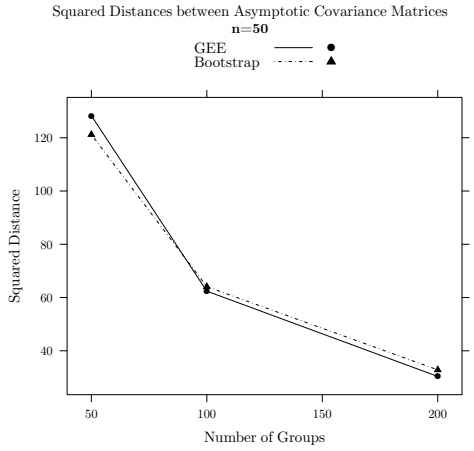
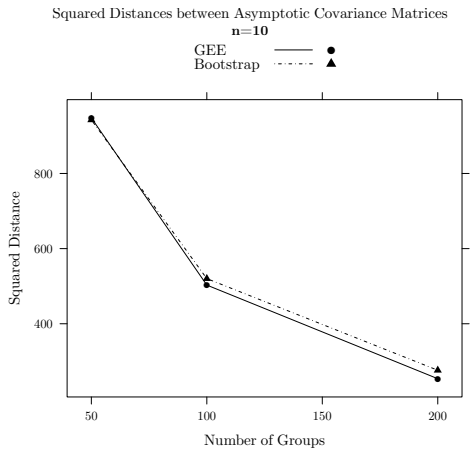
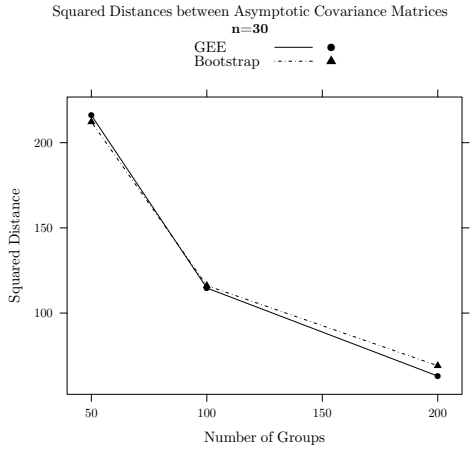


Figure 6.19: D^2 plots: asymptotic covariance matrices, $df = 9$ $ICC = .10$

(a)



(b)



(c)

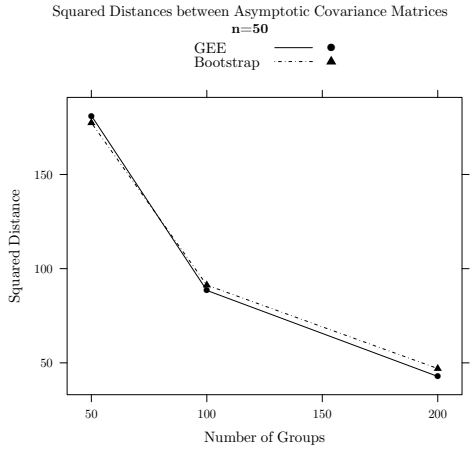
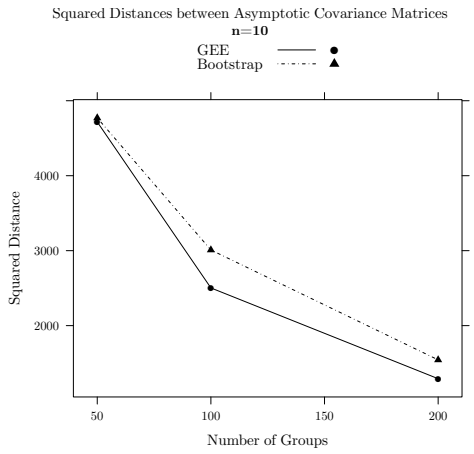
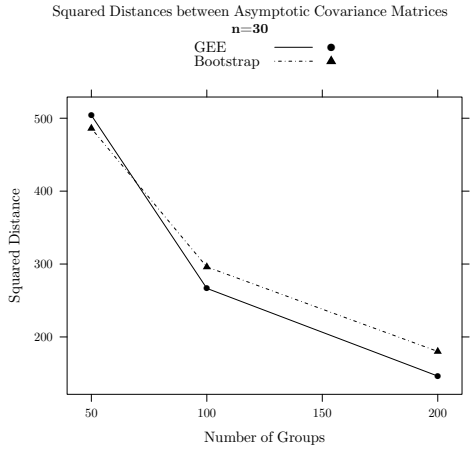


Figure 6.20: D^2 plots: asymptotic covariance matrices, $df = 9$ $ICC = .05$

(a)



(b)



(c)

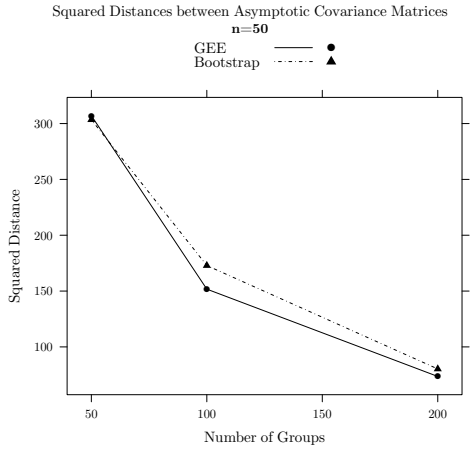
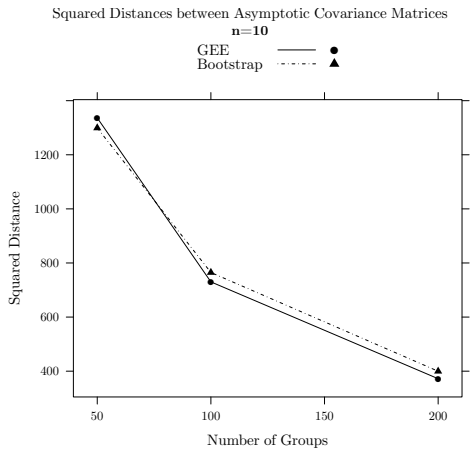
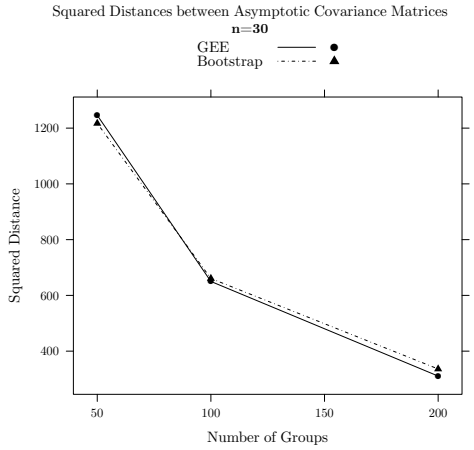


Figure 6.21: D^2 plots: asymptotic covariance matrices, $df = 54$ $ICC = .50$

(a)



(b)



(c)

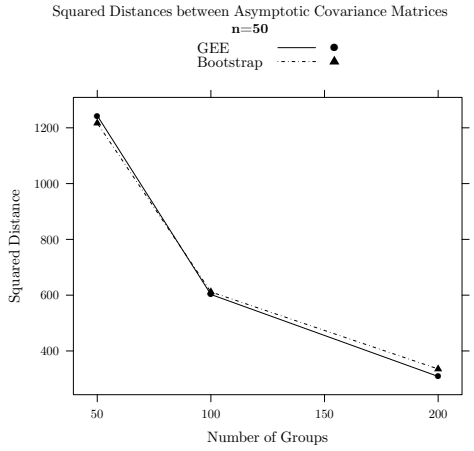
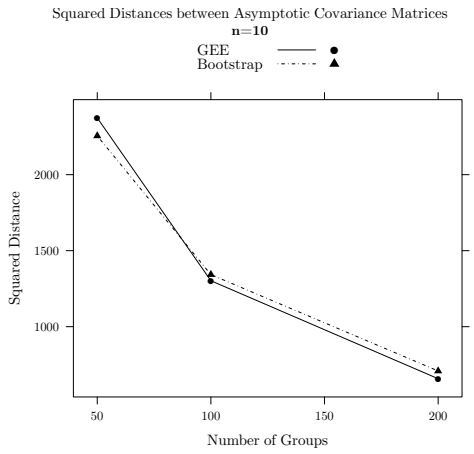
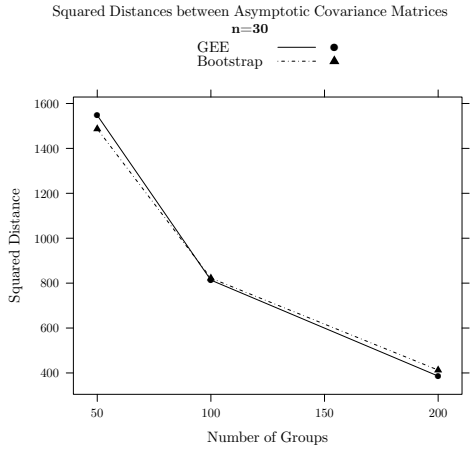


Figure 6.22: D^2 plots: asymptotic covariance matrices, $df = 54$ $ICC = .26$

(a)



(b)



(c)

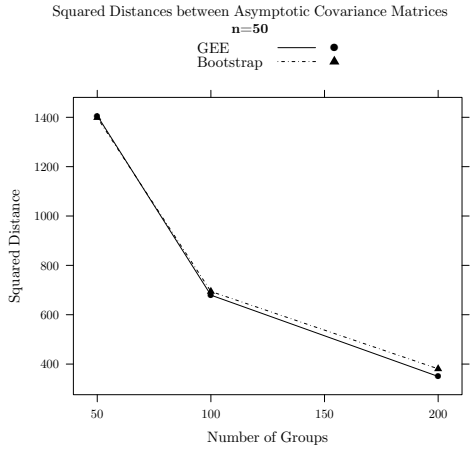
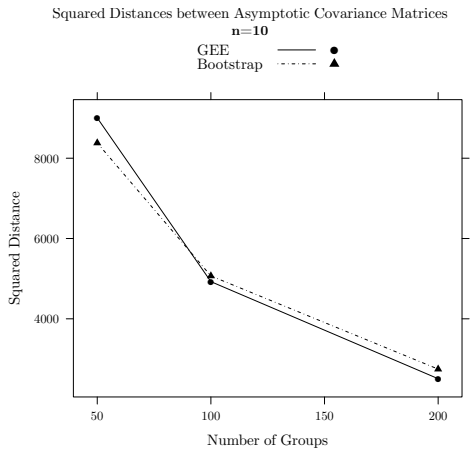
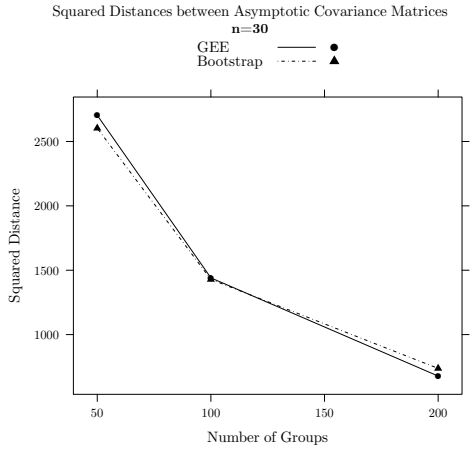


Figure 6.23: D^2 plots: asymptotic covariance matrices, $df = 54$ $ICC = .10$

(a)



(b)



(c)

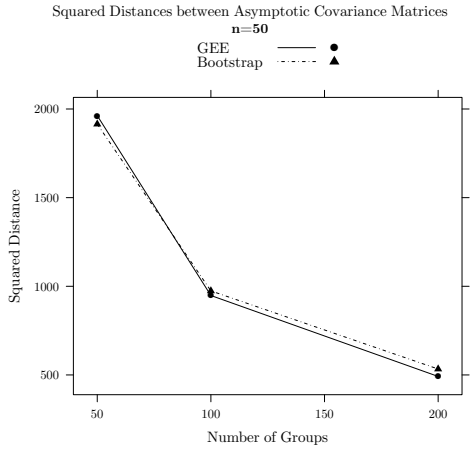
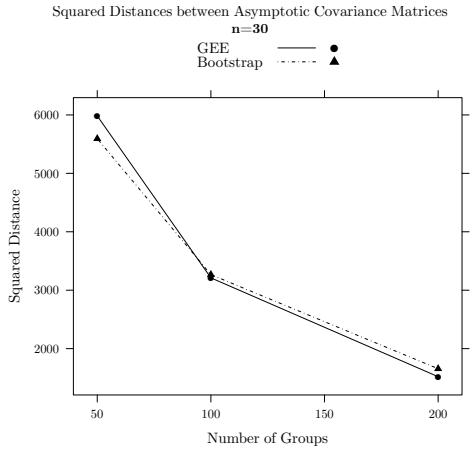
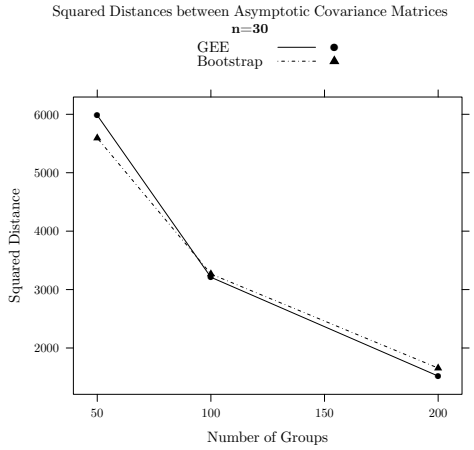


Figure 6.24: D^2 plots: asymptotic covariance matrices, $df = 54$ $ICC = .05$

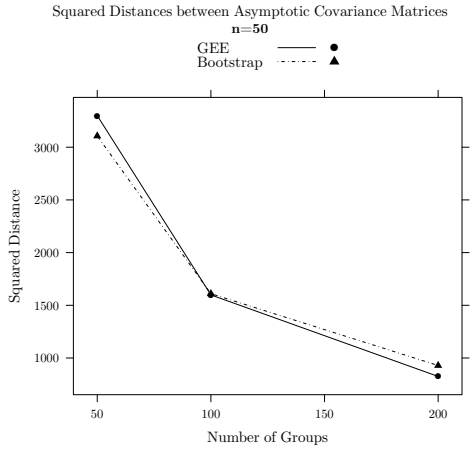
(a)



(b)



(c)



for the GEE-based approach with only 50 groups. This is consistent with past research findings (e.g., Feng et al., 1996; Mancl & DeRouen, 2001; Sherman & Cessie, 1997). However, this does not translate into better behaved test statistics with small level-2 sample sizes. In other words, for small samples, regardless of the precision of the estimation of Γ, test statistics are not properly distributed as central chi-square variates.

6.3 Summary of findings for simulation study 1

This section summarizes the results of Simulation study 1. Simulation study 1 examined the relative efficiency of the segregating approach, as compared to the partially saturated modeling method described in Section 3.5 (research topic 1), and examined the performance of ADF estimators based on $\hat{\Gamma}_{BOOT}$ and $\hat{\Gamma}_{GEE}$ under conditions likely to be encountered in realistic settings with student survey data (research topic 2). Simulation study 1 focused on two-level models—which would apply to situations where, for example, students were nested within classrooms. In addressing these questions, Simulation study 1 suggested four conclusions about the application of the segregating approach to the analysis of $\hat{\Sigma}_B$, the between-groups covariance matrix:

1. When used in conjunction with ML, parameter estimation using the segregating method is as efficient as the partially saturated model method under a wide range of conditions. At some low ICC conditions, the segregating method is relatively more efficient.

2. Maximum likelihood and ADF estimators both provide consistent estimates of model parameters. However, the ML estimator consistently shows lower mean-square error than either the bootstrap or GEE-based ADF estimators. Thus, it is recommended that ML estimation be used to obtain parameter estimates in conjunction with the segregating approach.

3. Standard errors based either on bootstrap or GEE approaches converged appropriately. However, the robust standard errors, based on the sandwich estimator, showed better performance, as measured by distance from the correct specified standard error at all sample sizes. Thus, robust standard errors are recommended. ML standard errors did not show appropriate convergence, particularly for low ICC conditions.

4. Bootstrap-based test statistics showed good performance for large level-2 sample sizes. It is recommended that the corrected residual-based test statistic T_{CRADF} be used, as this statistic showed the best performance across conditions for moderate to large sample sizes. For very large sample sizes, T_{RADF} also provides valid inferences about models. For small sample sizes, no test statistics show adequate performance. The likelihood ratio test statistic T_{ML} performed poorly overall. T_{RML} rescales the likelihood ratio test statistic, and always performs better than T_{ML}, but cannot adequately control Type I errors when ICCs are low or within group sample sizes are small.

6.4 Extension to three level models

The vast majority of writing about multilevel factor analysis considers only explicitly cases where there are two levels of nesting in the hierarchical structure: Persons nested within groups. There is little guidance on how to handle additional levels of hierarchy. For example, in the commonly encountered case of three level hierarchical data structures, with persons nested in subgroups nested in groups. While several key papers suggest that extensions to multiple levels is conceptually straightforward (Longford & Muthén, 1992; Yuan & Bentler, 2007), all but one of the simulation studies (Yau et al., 1993) provide empirical evidence for the applicability of the segregated method with two levels of nesting only. This dissertation

contributes to the literature by offering empirical evidence of the performance of the segregated approach in three level hierarchically structured data. Specifically, simulation study 2 was used to investigate whether the cluster bootstrap could be applied to three level hierarchical data sets in order to obtain estimates of Γ_{WG}, the between-subgroups level covariance matrix.

The results for simulation study 1 (section 6.3) were used to inform the selection of estimators and test statistics in simulation study 2. Specifically, the results of simulation study 1 showed that maximum likelihood estimates were consistent and unbiased, and had lower sampling variability than the ADF estimates in all simulation conditions (2 in section 6.3). Additionally, simulation study 1 showed that the cluster bootstrap based estimate of Γ, the asymptotic covariance matrix, could be used to estimate robust standard errors, and that those robust standard errors were consistent (3 in section 6.3). Finally, simulation study 1 showed that for adequate level 2 sample sizes, the test statistics T_{RADF}, T_{CRADF} and T_{RML} could potentially provide valid model inferences (4 in section 6.3).

Data were generated from multivariate normal distributions and a population model with one level-1 factor, one level-2 factor, and one level-3 factor (Figure 5.2). Each simulation condition consisted of 250 replications. Simulations were conducted using MPluss Monte Carlo capabilities. For each of the replicated data sets, MPlus and the MPlusAutomation package in R were used to obtain an estimate of Σ_{WG}. Model parameters and test statistics were estimated in EQS using the REQS package in R. The correct model was fit to each simulated data set.

6.4.1 Using the cluster based bootstrap with three level data

6.4.1.1 Accuracy of parameter estimates

Figure 6.25 present plots of the parameter bias and mean square error for the estimated factor loadings on the between-subgroup model of the simulated three level hierarchical data. The left panel displays the parameter bias, and the right panel displays the mean square error. As in Figures 6.1-6.4, parameter bias is plotted separately for each factor loading (λ). Unbiased factor loadings would have plotted values close to 0 (the center of the plot). Mean square error (MSE) is also plotted separately for each factor loading and, as in Figures 6.3-6.4, estimates are overall more accurate if MSE is close to 0 (the left margin of the plot). Figure 6.25 shows that the parameter bias is essentially 0 for all parameters, and there is very little variability in the parameter estimates, as the mean square error is close to 0. Recall that maximum likelihood estimation was used to obtain parameter estimates, and that ML estimation is not dependent on the use of the cluster bootstrap. These results give empirical evidence that the segregating approach can provide accurate and unbiased parameter estimates in three level models, just as it did for two-level models.

6.4.1.2 Test statistic distributions

The between subgroup (level-2) model in simulation study 2 has 9 degrees of freedom, equivalent to the small model condition in simulation study 1. Under the null hypothesis, the theoretical chi-square distribution has a mean of 9 and a standard deviation of $\sqrt{18} \approx 4.24$. Thus, the empirical mean and standard deviation should be close to this value in order for the statistic to be considered well-behaved. The empirical Type I error rates for the three test statistics can be compared to the nominal rate of $\alpha = .05$. As in simulation study 1, an acceptable empirical error rate is taken as one that falls in the interval [.020, .095], the

Figure 6.25: Parameter mean square error: three level model $df = 9$

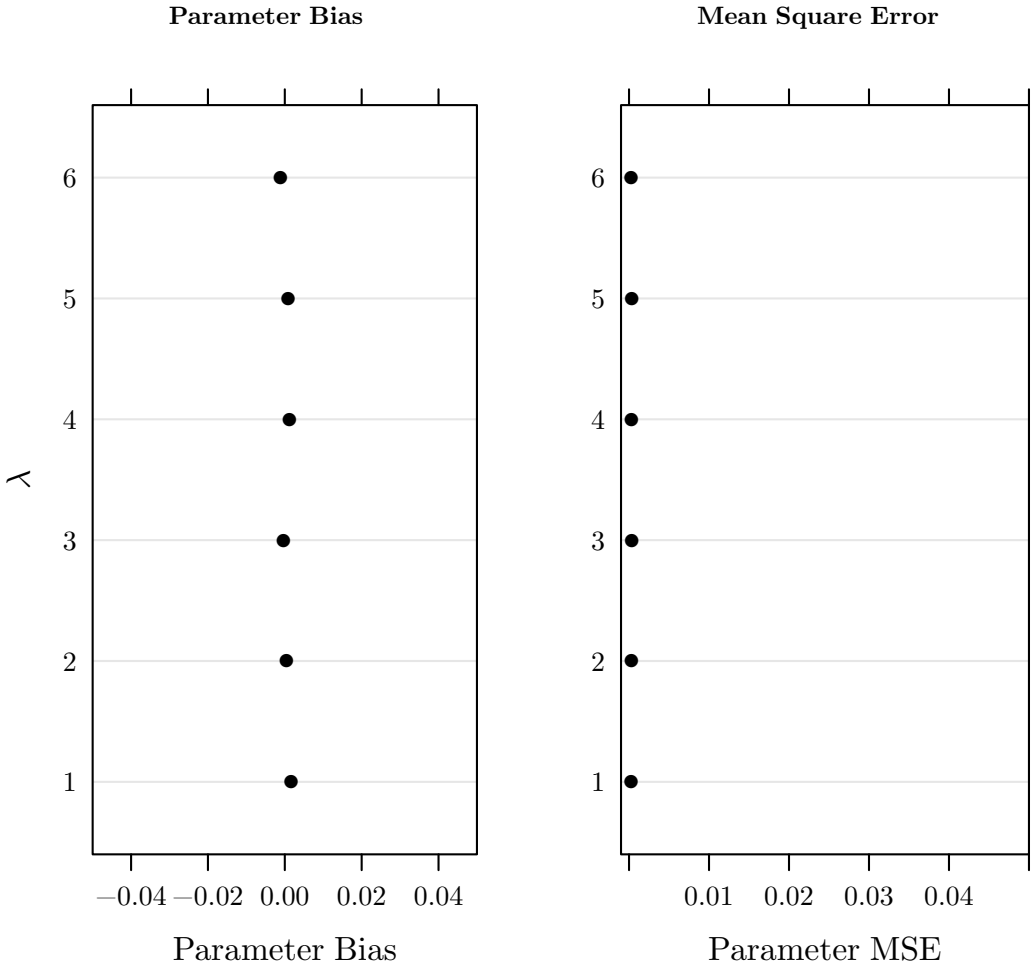
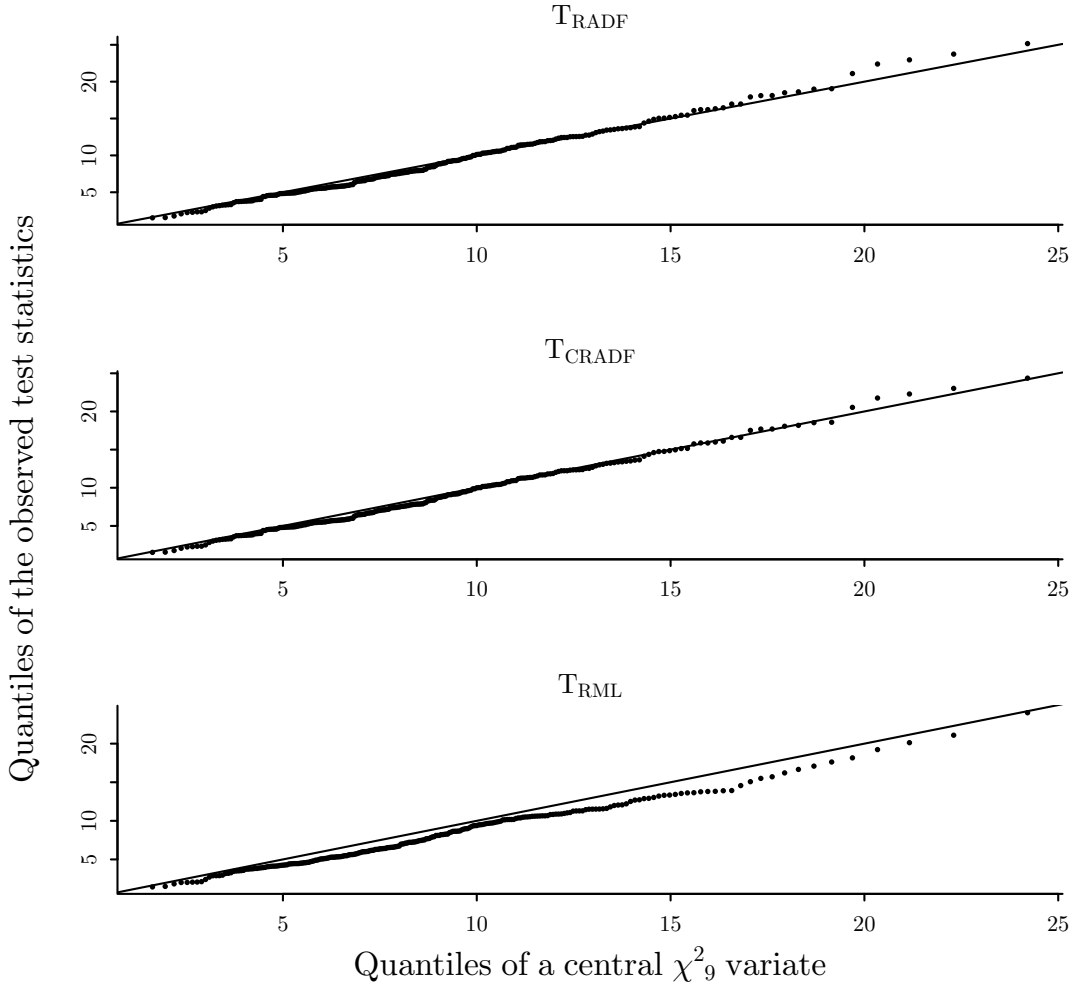


Figure 6.26: Q-Q plots: three level model $df = 9$



estimated 2-sided 99% adjusted Wald confidence interval (Agresti & Coull, 1998).

The residual based test statistics are well behaved, with means and standard deviations that are close to the theoretical values. Specifically, T_{RADF} has a mean of 8.81 with a standard deviation of 4.43, T_{CRADF} has a mean of 8.69 with a standard deviation of 4.32. The mean and standard deviation for the rescaled test statistic (T_{RML}) is actually slightly too small in application to three level models. T_{RML} has a mean of 7.96 with a standard deviation of 3.88. In terms of Type I error rates, the empirical Type I error rates are all within range of the .05 nominal value. T_{RADF} has an empirical Type I error rate of .056, T_{CRADF} has an empirical

Type I error rate of .048, and T_{RML} has an empirical Type I error rate of .028. Figure 6.26 presents Q-Q plots for the three test statistics. The solid line in each plot represents the theoretical distribution, and the empirical distribution more closely follows the theorized chi-square distribution if the points follow this line closely. The Q-Q plots show that T_{RADF} and T_{CRADF} are well behaved and match expectation closely, even in the tails. T_{RML} is systematically lower than the theoretical value, and this discrepancy is even more pronounced in the upper tail.

6.5 Summary of findings for simulation study 2

This section summarizes the results of simulation study 2. Simulation study 2 investigated whether a cluster bootstrap-based approach to estimating the between-group asymptotic covariance matrix be extended to data sets with three levels. Specifically, simulation study 2 examined the use of maximum likelihood estimation to obtain parameter estimates, and the use of $\hat{\Gamma}_{BOOT}$ to estimate robust standard errors and residual based and rescaled test statistics, T_{RADF}, T_{CRADF} and T_{RML}. Overall, simulation study 2 suggests the following conclusions:

1. ML estimation provides precise and unbiased estimates of model parameters in three level models. Thus, as was the case in two-level models, it is recommended that ML estimation be used in conjunction with the segregating approach to obtain parameter estimates.

2. Residual-based test statistics using $\hat{\Gamma}_{BOOT}$ showed good performance for large sample sizes in three level models. For sufficient sample sizes, it is recommended that T_{RADF}, and T_{CRADF} be used, as these statistics showed the best overall performance. The rescaled test statistic T_{RML} had a mean and standard deviation that were slightly too low for use in three level models, but the Type I error rate was acceptable.

While simulation study 2 was significantly smaller in scope than simulation study 1, it makes a valuable contribution to the research on multilevel factor analysis using the segregated approach. The cluster bootstrap has never been applied to multilevel factor analysis, and while the theory of the bootstrap would suggest that such applications are possible (e.g., Ren et al., 2010), an important first step is to establish empirically that the theory holds under realistic data conditions (e.g., Yung & Bentler, 1994). Simulation study 2, thus, offers evidence that, given adequate sample sizes, the cluster bootstrap can be applied to datasets with multiple levels of nesting. This establishes the foundation for Chapter 7, which offers an illustration of the segregating method on a realistic dataset to investigate the dimensions of teacher quality that are discernible in a state-wide student survey of instructional practice.

CHAPTER 7

Empirical Illustration of a Three Level Factor Analysis using the Cluster Bootstrap: New Mexico Student Survey

The fourth research topic concerned the application of the cluster bootstrap methods to a realistic dataset to investigate the dimensions of instructional practice that are discernible in a state-wide student survey of instructional practice. First, descriptive information about the dataset is presented. Second, the dimensionality of the survey is explored. Finally, the relationship between the survey constructs and student achievement is investigated.

The Opportunity to Learn (OTL) Survey is a 10 item survey designed to measure the quality of instruction and the school environment. Different versions of the survey are administered in elementary (grades 3-5) and middle and high school (grades 6-12). Each item is scored on a 6-point scale, from 0 to 5, where the categories are 0 = *never*, 1 = *hardly ever*, 2 = *sometimes*, 3 = *usually*, 4 = *almost always*, and 5 = *always*.

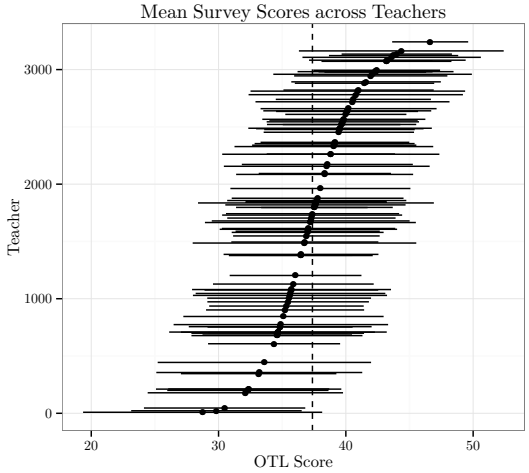
Data used in this study were collected in the 2012-2013 administration of the OTL survey. This analysis focuses only on the early grades version of the survey, where student raters are uniquely nested within a single teacher. Table 7.1 shows descriptive statistics for each item on the survey, including the mean, standard deviation, and the amount of total item variance that is accounted for by classroom and school levels, respectively.

Table 7.1: Item descriptives: OTL Survey

Item	Mean	Sd	variance (teacher)	variance (school)
My teacher introduces a new lesson by reminding us of things we already know.	3.48	1.33	7.89%	3.70%
My teacher explains why what we are learning is important.	3.82	1.26	5.28%	2.73%
My teacher explains how learning each lesson will help us in the future.	3.43	1.45	7.32%	2.18%
Everybody gets a chance to answer questions.	3.66	1.4	6.45%	1.84%
My teacher wants me to explain my answers.	4.09	1.23	8.15%	3.07%
My teacher explains things in different ways so everyone can understand.	3.85	1.31	5.86%	2.06%
My teacher helps me when I do not understand.	4.07	1.28	10.12%	2.79%
I use different materials and tools to help me practice what I am learning.	3.14	1.4	6.41%	2.95%
My teacher makes sure I understand.	4.5	0.96	4.55%	1.14%
My teacher take the time to summarize what we learned each day.	3.28	1.51	7.41%	3.05%

In general, the mean responses show fairly positive ratings, with three items having overall means above 4 on a 0-5 scale. This suggests that students feel relatively positively about the instructional quality in their classroom. For each item, the percentage of variance at the classroom level is greater than the percentage of variance at the school level. This suggests that there is more variation in student ratings across classrooms than there is across schools. This is consistent with past research on teacher quality, which has found greater variability within schools than across schools (Rowan & Correnti, 2009). This decomposition also suggests that nearly 90% of the item variance is between students within the same classroom—i.e., it is not explained by factors at either the teacher or the school level. In order to visualize this, consider Figure 7.1, which shows the mean OTL scores for a random sample of 100 teachers across all schools. Classroom means are represented by the dots (•), and the horizontal bars extend ±1 standard deviation. From this plot, it is clear that, while there is variation in classroom mean scores, there is tremendous variability across-students within classrooms.

Figure 7.1: Distribution of class-mean survey scores



7.0.1 What dimensions of instructional practice are discernible aggregated student responses in the OTL survey?

In order to determine the dimensions of instructional practice that are discernible based on aggregated student responses in the OTL Survey, a multilevel factor analysis using the segregating method was used to following the method outlined and tested in the previous chapters. Specifically, since there is a three level hierarchical structure to the data set, with students nested in classrooms nested in schools, the segregating method was used to first extract the between classroom covariance matrix. A cluster bootstrap was used to obtain an estimate of the asymptotic covariance matrix by resampling intact schools. Maximum likelihood was used to estimate parameters, and robust standard errors were used to make inferences about those parameters. The residual-based test statistics T_{RADF} and T_{CRADF} were used for model appraisal. In addition, the Root Mean Square Error of Approximation (RMSEA) (Steiger & Lind, 1980), the Comparative Fit Index (CFI) (Bentler, 1990) and the average absolute standardized residual were inspected. The between-teacher covariance matrix is given in the first ten columns of Table 7.2. There are two possible conceptual models behind the OTL survey. On one hand,

Table 7.2: Estimated sample between-teacher covariance matrix

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	VAM_R	VAM_M
Q1	0.14											
Q2	0.07	0.08										
Q3	0.12	0.10	0.16									
Q4	0.08	0.06	0.08	0.13								
Q5	0.08	0.04	0.07	0.05	0.12							
Q6	0.09	0.06	0.09	0.07	0.06	0.10						
Q7	0.09	0.04	0.09	0.06	0.07	0.09	0.16					
Q8	0.09	0.07	0.09	0.08	0.05	0.07	0.05	0.12				
Q9	0.05	0.03	0.05	0.04	0.03	0.05	0.06	0.03	0.04			
Q10	0.09	0.08	0.10	0.08	0.05	0.07	0.05	0.09	0.03	0.17		
VAM_R	0.01	0.01	0.00	0.01	0.01	0.00	-0.01	0.01	0.00	0.01	0.06	
VAM_M	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.03	0.04

it is possible to conceive of instructional practice as being unidimensional. On the other hand, it is possible that there are discernible subdimensions of instructional practice, that explain covariation among the items in addition to the general factor. These subdimensions could include, for example, "social support" (the extent to which a teacher creates a supportive social environment, as described in Kunter et al. (2008)) and "context" (the extent to which teacher provide students with a sense of instructional trajectory, as described in Patrick et al. (2003)).

Because there are two different conceptual models, both a unidimensional model and a bifactor model with one general factor and two specific factors were fit to $\hat{\Sigma}_{WG}$, the estimated between-teacher covariance matrix. The unidimensional model would support a theory that there is one overall instructional practice factor. The bifactor model would support a theory that there are two domain specific factors, *social support* and *context*, each of which account for unique variance beyond the general instructional practice factor. Standardized parameter estimates for each model are reported in Table 7.3. It should be noted that, regardless of which test statistic is used (i.e., T_{RADF} or T_{CRADF}), the null hypothesis that the population covariance matrix is a function of the model parameters is rejected by the chi-square test for both the unidimensional and the bifactor models. These test statistics can be compared to a central χ^2_{35} distribution in the case of the

Table 7.3: Standardized factor loadings: OTL survey

	Unidimensional		Bifactor			
	Loading	Uniqueness	Loadings			Uniqueness
Item	General	Uniqueness	General	Context	Social Support	Uniqueness
1	.892	.453	.915			.402
2	.802	.597	.779	.524		.345
3	.873	.487	.855	.379		.353
4	.694	.720	.702			.712
5	.637	.771	.643			.766
6	.872	.489	.835		.293	.465
7	.713	.701	.647		.575	.500
8	.725	.689	.757			.654
9	.777	.629	.701		.712	.033
10	.671	.741	.689			.725

All parameters significant at the $\alpha = .05$ level.

Table 7.4: Fit statistics and fit indices for unidimensional and bifactor models

	Model	
	Unidimensional	Bifactor
T_{RADF}	495.22	278.29
T_{CRADF}	421.54	253.40
RMSEA (T_{RADF})	.068	.053
RMSEA (T_{CRADF})	.062	.050
CFI (T_{RADF})	.46	.53
CFI (T_{CRADF})	.69	.75
ASR[a]	.054	.035

Note:[a]ASR=Absolute Standardized Residual

unidimensional model, and a central χ^2_{31} distribution in the case of the bifactor model. This indicates a non-negligible amount of misfit between the data and the models. However, based on fit indices including RMSEA, CFI and the average absolute standardized residuals, the bifactor model is a better fit (Table 7.4). In particular, the CFI is higher for the bifactor model, and the RMSEA and absolute standardized residuals are smaller. Thus, there is more support in the data for the theory that, in addition to being able to discern overall instructional practice, aggregated student responses can be used to distinguish two domain specific factors— *social support* and *context*—each of which account for unique variance beyond the general instructional practice factor.

7.0.2 How do these survey-derived variables relate to outcomes of policy interest, such as student achievement gains?

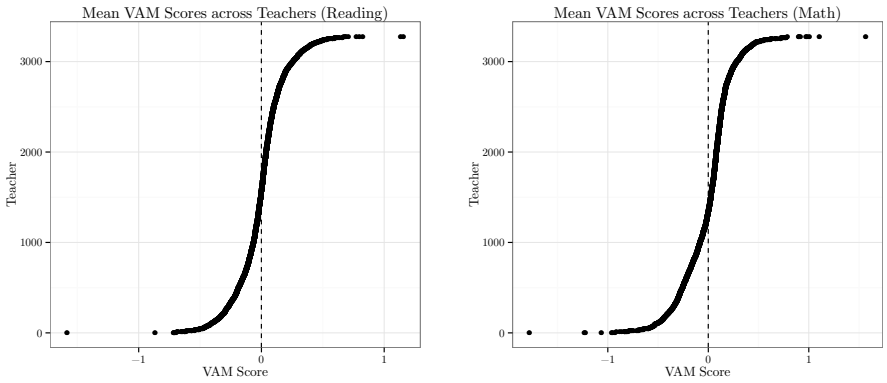
Since the bifactor model was preferred over the unidimensional model, there is support for the theory that the aggregated survey responses could be explained by one general instructional practice factor and two specific factors, defined as *social support* and *context*. It is now possible to illustrate how the segregating approach can be used to explore relationships between latent variables and external variables of policy interest. Specifically, it is possible to examine whether aggregated ratings of instructional practice predict student achievement gains in math and reading, and whether the specific factors of context and social support are predictive "over and above the general factor" (Chen et al., 2006, p. 197).

The VAM scores do not have any within-classroom variation, and so the steps taken in this analysis follow the general model outlined in Section 3.7, and given in Equation 3.59. As was described in Section 3.7, even when variables with level restricted variation are included, estimators such as those given in Equation 3.54 can still be used to obtain consistent estimates of the between-classroom level covariance matrix. The variance and covariances of the two VAM scores (labeled as VAM_M for math and VAM_R for reading) are reported in Table 7.2.

The VAM_R estimates have a mean of 0 and a standard deviation of .20, and the VAM_M estimates have a mean of 0 and a standard deviation of .25 (math). Approximately 90% of the variance in the VAM scores is between teachers within schools (10% is between schools). Figure 7.2 shows that there are a great number of teachers close to the mean (indicated by the vertical dotted line), and a few teachers that are well above or below this mean.

Model parameter estimates are reported in Table 7.5. The general factor (instructional practice) significantly predicts VAM scores for both math and reading. This offers some important validity evidence for making inferences about professional practice based on the OTL survey, as the relationships between professional prac-

Figure 7.2: Distribution of VAM scores



(a) VAM_R (b) VAM_M

Table 7.5: OTL Survey and External Criterion: VAM Math

Item	General	Context	Social Support	Uniqueness
1	0.915			0.404
2	0.769	0.540		0.344
3	0.851	0.393		0.345
4	0.701			0.713
5	0.634			0.771
6	0.827		0.299	0.476
7	0.647		0.563	0.510
8	0.750			0.662
9	0.717		0.695	0.000[a]
10	0.691			0.725
VAM_R	0.092	0.057[a]	-0.047[a]	0.993
VAM_M	0.172	0.081[a]	-0.003[a]	0.982

$T_{RADF} = 526.45$. $T_{CRADF} = 443.95$ Test statistics can be referred to χ^2_{46}.
All parameters significant at the $\alpha = .05$ level except those marked as [a].

tice and VAM scores were positive. However, the specific factors do not predict VAM scores beyond the general factor for either math or reading.

CHAPTER 8

Summary and Discussion

Student surveys continue to be one of the most popular and widespread mechanisms for collecting information about instructional practice. Advocates of using student surveys for either formative or summative evaluation note that students are natural observers of their classroom environments, and that student ratings of teacher practice show relatively robust correlations with student achievement (Burniske & Meibaum, 2011). In a 2012 speech at the Education Commission of the States National Forum on Education Policy, Bill Gates noted that the three components of a good evaluation system are test scores, observations, and student ratings (Gates, 2012).

However, even as their application has increased, there has been relatively little research into the psychometric properties of these surveys. This is particularly true of *multilevel* psychometric research, investigating the psychometric properties of the aggregated student ratings.

This dissertation investigated four research topics with the objective of addressing several open issues in multilevel factor analysis, and focused specifically on issues that were unaddressed in the research literature in the application of multilevel factor analysis models to aggregated student ratings of instructional practice. Specifically, this dissertation examined the performance of the segregating approach to multilevel factor analysis under real world conditions in order to study: 1) The efficiency of the segregating approach compared to the partially saturated model method in the estimation of parameters in two-level models. 2) The com-

parative performance of GEE-based ADF, cluster bootstrap-based ADF and ML estimators in the segregated analysis of between-groups covariance structure 3) The extension of the bootstrap-based method to three level models. 4) the extension of the cluster bootstrap-based method to three level models. 4) The application of these bootstrap methods to a realistic dataset to illustrate how the methods may be used to investigate the dimensions of teacher professional practice that are discernible in a state-wide student survey of instructional quality. The results are worth consolidating and discussing in more detail here.

8.1 The segregating approach is relatively efficient

The segregating approach is implemented in two steps, and so there are questions as to whether the loss of efficiency that arises from the use of a two-step method is substantial enough to dissuade researchers from implementing the method, and using a one-step method of model evaluation—such as the partially saturated modeling method—instead. This dissertation demonstrated that, when evaluating the estimation of group level parameters, for balanced group sizes and two level models, using maximum likelihood estimation in conjunction with the segregating approach yields parameter estimates that are at least as efficient as those from the partially saturated modeling method. Beyond that, this dissertation offered empirical evidence to support the hypothesis of Yuan and Bentler (2007) that in some cases, the segregating approach may be far more efficient than the partially saturated modeling method. While Yuan and Bentler (2007) acknowledged that this could happen as model complexity increases, this dissertation demonstrated that the segregating approach is relatively more efficient for small models as within-group sample sizes or ICCs decrease. The relative efficiency of the segregating approach, along with the fact that it is readily implemented into virtually any conventional

factor analysis or covariance structure analysis software, makes the segregating approach an appealing approach to researchers studying the psychometrics of student surveys of instructional practice, where those exact conditions—small within-group sample sizes and low ICCs—are commonly encountered.

8.2 The cluster bootstrap can be used to obtain test statistics and standard errors, provided sample sizes are sufficient

Because maximum likelihood estimation produces the most accurate and precise parameter estimates, it is recommended to use maximum likelihood estimation in conjunction with the segregating approach, particularly for the analysis of the level-2 covariance matrix. However, it is not recommended that the widely used Likelihood Ratio test statistic T_{ML} be used to test the null hypothesis of exact fit, because this statistic will not be asymptotically distributed as a central chi-square variate when analyzing the between-groups covariance matrix. Additionally, maximum likelihood estimated standard errors tend to be too small, even when the model is correctly specified—particularly when ICCs are low or within-group sample sizes are small.

Results in this dissertation suggest that the residual-based test statistics T_{RADF} and T_{CRADF} be used for model testing. These statistics, and in particular T_{CRADF}, showed the best performance over a wide variety of simulation conditions. These statistics are asymptotically distribution free and can be obtained in the framework of ML estimation. In addition, it is recommended that the robust (sandwich estimated) standard errors be used for inferences about model parameters, as those standard errors showed the best performance over a wide variety of simulation conditions, and particularly in cases when item ICCs are low or within-group sample sizes are small.

The residual-based test statistics and robust standard errors require an estimate of Γ_B, the asymptotic covariance matrix. This dissertation demonstrated that consistent estimates of this asymptotic covariance matrix can be obtained using the cluster bootstrap. As far as this author knows, this is the first time that the cluster bootstrap has been applied to the problem of estimating an asymptotic covariance matrix for use in conjunction with the segregating method. The cluster bootstrap has many possible advantages over the Generalized Estimating Equation (GEE) framework proposed in Yuan and Bentler (2007). Specifically, while the GEE framework requires researchers to specify a matrix of partial first order derivatives of the log likelihood function. And, while Yuan and Bentler (2007) provided the relevant mathematical details (and a SAS macro to do the computation), the specification of this matrix becomes increasingly more complicated as the number of hierarchical levels increases. For example, details on the relevant log-likelihood function for three or four level models are not readily available.

On the other hand, the non-parametric cluster bootstrap requires no such specifications—researchers only need access to software capable of performing the relevant re-sampling (with replacement). This is far less technically and mathematically challenging, and much more straightforward to implement. In fact, it could be argued that these benefits of the cluster bootstrap outweigh some of the negatives—specifically, that the cluster bootstrap requires larger sample sizes than GEE estimation for test statistics to be properly distributed, and the cluster bootstrap is more computationally expensive, in the sense that it will take longer to run, even on relatively high performance computer systems.

8.3 The cluster bootstrap can be extended to three level models

In many real world applications, and in particular in the context of student surveys of instructional practice or teacher quality, it is necessary to consider hierarchical data structures with multiple levels of nesting. For example, a three level model with students nested within teachers, and teachers nested within schools. In fact, in the context of secondary school teachers, where individual teachers may teach multiple sections of a class, it may be necessary to consider hierarchical models with four levels: students nested in class sections, nested in teachers, nested in schools.

In spite of this, the research base on how best to implement factor analytic methods into data structures with more than two levels of nesting is very small. While Yuan and Bentler (2007) and Longford and Muthén (1992) noted that the theoretical basis of multilevel factor analysis is readily expanded from two level setting to settings with two and three levels, and Hox (2010) and Goldstein (2003) noted that obtaining estimates of the relevant covariance matrices for use in the segregating method is easily done using conventional software, only Yau et al. (1993) investigated with an empirical example how the segregating approach may be used in multilevel data sets with three levels of nesting. However, their examples use conditions that are not likely encountered in the context of student surveys of professional practice or instructional quality.

Thus, this dissertation makes three contributions to the research on multilevel factor analysis. Through a simulation study, it was demonstrated that the segregating approach can be used in conjunction with maximum likelihood estimation in data sets with three levels of nesting to yield accurate parameter estimates. The simulation study also demonstrated that the cluster bootstrap could be extended to three levels to estimate Γ_{WG}, and that—for adequate level 2 sample sizes—the

residual-based test statistics T_{RADF} and T_{CRADF} based on this estimate will be appropriately distributed and can be used to make inferences about measurement models. Finally, this dissertation illustrated how the segregated approach can be applied to a realistic student survey dataset, and illustrated how the segregated approach can be used to test measurement models and to explore relationships between aggregated survey variables and outcomes of importance to policy and practice.

8.4 Limitations of the current study

This dissertation relied on a series of simulation studies in order to make inferences about the applicability of the segregating approach to a range of conditions and data configurations. However, as with any simulation study, caution should be used in generalizing these results to other conditions not included in the study. There are several conditions that were not included in this dissertation, in particular, that are worth mentioning here, as more work is needed to investigate how these conditions would influence parameter estimation and test statistic performance.

8.4.1 Non-normal distributions

All of the population models used in the simulation studies generated data that was multivariate normal in distribution. However real data is rarely normal in distribution (Micceri, 1989), and observed variables typically exhibit excess skew and kurtosis, relative to a normal distribution. When indicators are excessively skewed or kurtotic, there is limited research in multilevel factor analysis using the segregating method demonstrating empirically that the residual based test statistics—which are based on asymptotic distribution free theory—are appropriately distributed under the null hypothesis (e.g. Bentler & Yuan, 1999). However,

there is substantial research on this issue in conventional factor analysis (e.g., Bentler & Yuan, 1999; Hu et al., 1992; B. O. Muthén & Kaplan, 1985, 1992).

Regarding standard errors, there is less research that examines how standard errors of parameter estimates—and in particular, the robust standard errors—perform under distributional violation in multilevel factor analysis or covariance structure analysis. Yuan and Bentler (2006) explored analytically the effect of skew and kurtosis on standard errors in multilevel factor analysis, and recommended the use of robust (sandwich estimated) standard errors, though the segregating approach in particular was not studied.

Finally, while in theory the cluster bootstrap is non-parametric and should provide consistent estimates of the asymptotic covariance matrix under a wide variety of distributions, the performance of the cluster bootstrap in this context has not been studied. The limited work that does exist on the cluster bootstrap (Carpenter et al., 1999; Field & Welsh, 2007; Ren et al., 2010; Samanta & Welsh, 2013; Van der Leeden et al., 1997) is focused on univariate models.

8.4.2 Simplified generating model

The population models used in simulation here were relatively simple. For example, in simulation study 1, the between-groups model contained only one factor, which influenced the observed indicators. In that study, observed scores for individuals were simulated based on a common factor model, and so the observed scores are exact linear combinations of the factor scores and the uniquenesses. However, as MacCallum and Tucker (1991) pointed out, in practice, this generating model is unrealistic, and there is some amount of model error that is unaccounted for. In other words, there is a "lack of correspondence between the model and the population covariance matrix" (p. 507). For example, Tucker, Koopman, and Linn (1969) noted that there may be many minor factors that are not of substantive interest or importance, but that influence the values of the observed scores, and

incorporated these minor factors into simulation models, to see how these would influence model appraisal.

8.4.3 Balanced group sizes

All of the simulation studies included in this dissertation used balanced groups. That is, all of the groups in simulation study 1, and all of the subgroups in simulation study 2 contained the same number of units. This condition is not likely to be encountered in real world research settings, where numbers of students vary from classroom to classroom, and numbers of teachers vary from school to school.

While results in Yuan and Bentler (2006, p. 2007) suggested that the ADF and residual-based test statistics used in this study should be correctly distributed in the segregating approach even as group sizes are unbalanced, there is little empirical work demonstrating how differences in group sizes influence test statistic performance and convergence. Additionally, the influence of unbalanced group sizes on parameter estimation, accuracy and efficiency have not been systematically studied. There is also little research on the influence of unbalanced group sizes on the performance of the cluster bootstrap, and the limited work that does exist on the cluster bootstrap focuses on univariate models.

8.5 Directions for future research

This study helped to develop and test a framework for implementing the segregating approach to three level factor models, and focused specifically on the psychometric properties of aggregated indicators of teacher quality and instructional practice. The results of this study suggest five areas for future research, each of which are briefly addressed here.

8.5.1 Crossed raters

In elementary school settings, the multilevel factor models developed and tested in this dissertation adequately describe the hierarchical structure of the data. That is, the model assumptions are that student raters are uniquely associated with teachers, and that teachers are uniquely associated with schools. Each student rates only one teacher, and each teacher appears in only one school. This is sensible for mainstream elementary school settings, because the typical model for elementary education is to have one general subject teacher providing instruction to a stable set of students in one school building for an entire academic year.

However, this model is less sensible when applied either to 1) special education or subject specialist teachers at an elementary level or 2) secondary school teachers. Special education teachers frequently move classrooms or move school buildings. Secondary school teachers are often subject specific, and it is common for students to have separate English, mathematics, and science teachers, say, in middle school and high school.

When the same student rates multiple teachers, or when teachers move from school to school, the assumption of a hierarchical data set is violated. This sort of model may be more accurately described as a cross-classified model. Little research currently exists on how to incorporate cross-classified rater effects into factor models or covariance structure models, and more research is needed in this area.

8.5.2 Measurement error or substantive variation: differences between students within classrooms

The measurement models used in this study are based on an assumption that variance between teachers within schools is substantively interesting, and represents meaningful differences in instructional quality, but that variance between students

within classrooms represents rater error, and should not be substantively interpreted (e.g., Marsh et al., 2012).

The tenability of the assumption that variance between students within a classroom is attributable to error is questionable. It is often difficult to distinguish items intended to measure individual, psychological constructs from items that are intended to measure organizational constructs, and many items are not readily categorized, and it is possible to conceive of microclimates, where individual students have legitimately different experiences with instruction in a particular classroom. This raises important questions about what is being measured by a particular survey. Are the items measuring qualities of the classroom? Or are they measuring qualities of the students? Or both? (e.g., Sirotnik, 1980). Alternative models that consider the possibility that differences in ratings across students within a classroom represent meaningful differences between students should be developed and explored.

8.5.3 Nonlinear latent variable modeling frameworks

The models in this dissertation are extensions of commonly used confirmatory factor analysis models, which attempt to structure the covariances between items. In general, these models are built upon an assumption that items are continuous, rather than categorical.

There exist other models for categorical data, that do not treat the observed indicators as continuous, including multilevel item response models (Fox & Glas, 2001) or multilevel item factor models, as well as other nonlinear latent variable models (Yang & Cai, 2012; Yang, Monroe, & Cai, 2012). There has been relatively little application of these models to student ratings of instructional practice, and this is an area that should be further developed and explored.

8.5.4 Comparison of aggregated and disaggregated analyses

While there is a long tradition of literature (e.g., Cronbach, 1976; Härnqvist, 1978; Julian, 2001; Longford & Muthén, 1992; Zyphur et al., 2008) demonstrating the advantages of multilevel factor anlaysis and suggesting that that single-level analytic methods that do not account for hierarchical data structures are problematic and can be "substantively misleading" (Reise et al., 2005, p. 130), they are still widely used in applied research. In the context of student ratings of professional practice, one relatively unexplored area of research concerns comparisons of the precision and accuracy of teacher scores that can be estimated using multilevel factor analytic techniques such as the segregating method compared to other methods. For example, two widely used analytic approaches involve either 1) assigning teacher scores based on the means of observed student scores (e.g., Mihaly et al., 2013), or 2) using a disaggregated factor analysis to compute factor scores for individual students, and then aggregating those factor scores to assign teacher scores (e.g. DiStefano et al., 2007). In the context of appraising teacher quality, when a single score is created from all of the survey items, there many open issues regarding how inferences about individual teachers change depending on the modeling approach, and whether (and how) perceptions of precision are influenced by model choice.

8.5.5 Small samples and large surveys

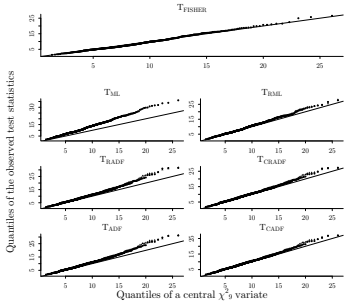
One last area of potential future research concerns the investigation of methods—whether covariance structure methods such as those investigated in this dissertation, or alternative methods such as nonlinear latent variable models or Bayesian latent variable models—when sample sizes are small and measurement models complex. The results of this dissertation suggest that for small level 2 sample sizes, parameter bias, efficiency, and test statistic performance makes model estimation and appraisal difficult. However, many studies involving student ratings of instructional practice

contain more items than the survey included in this dissertation. At the same time, the number of teachers, classrooms, or schools tend to be far smaller than were considered here. How best to estimate models and make inferences about the psychometric properties of surveys under these conditions is largely unknown.

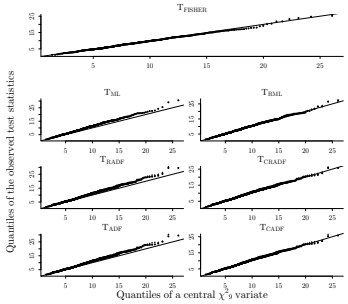
Appendix A

Additional Q-Q Plots

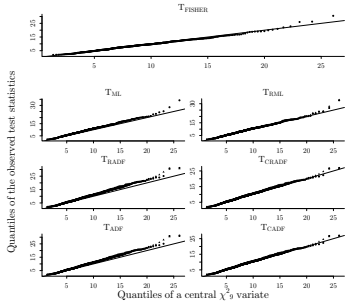
Figure A.1: Q-Q plots $df = 9$, $ICC = .50$



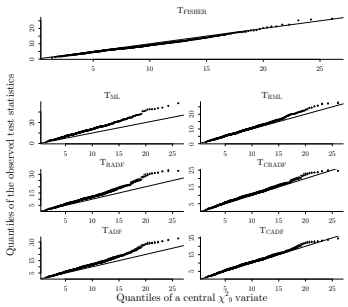
(a) $J = 200$ $n = 10$



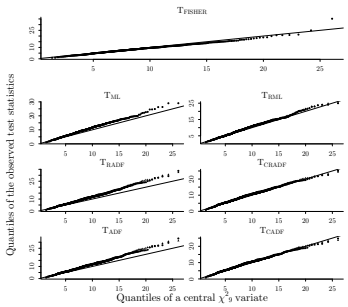
(b) $J = 200$ $n = 30$



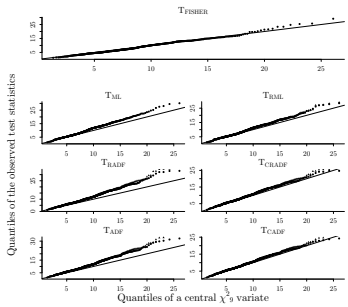
(c) $J = 200$ $n = 50$



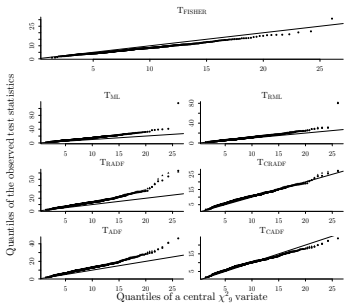
(d) $J = 100$ $n = 10$



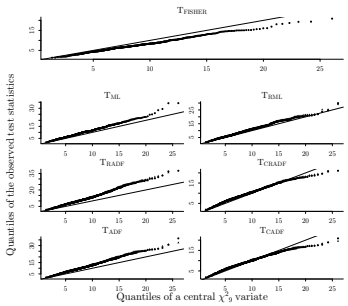
(e) $J = 100$ $n = 30$



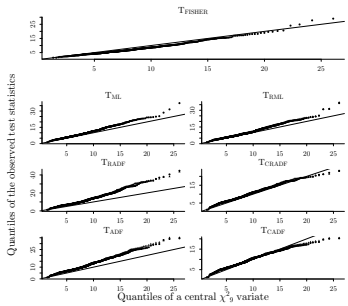
(f) $J = 100$ $n = 50$



(g) $J = 50$ $n = 10$



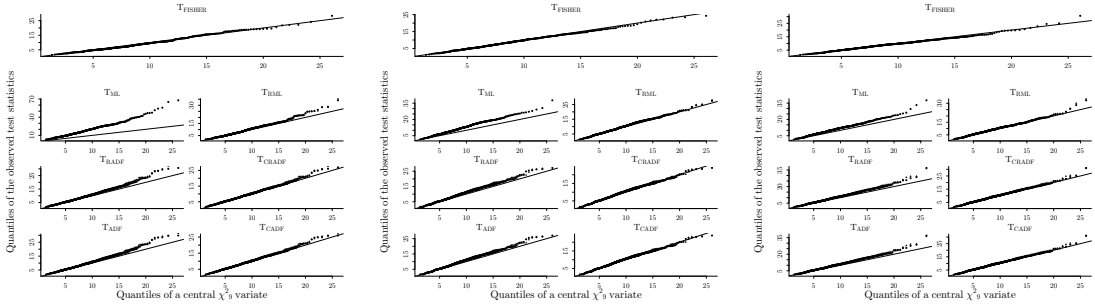
(h) $J = 50$ $n = 30$



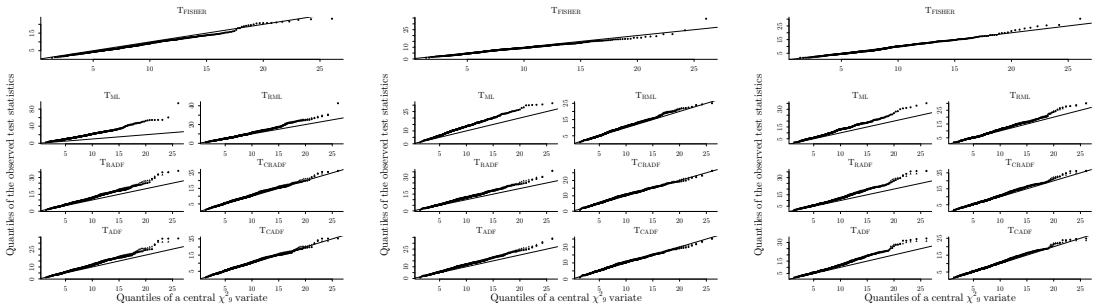
(i) $J = 50$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

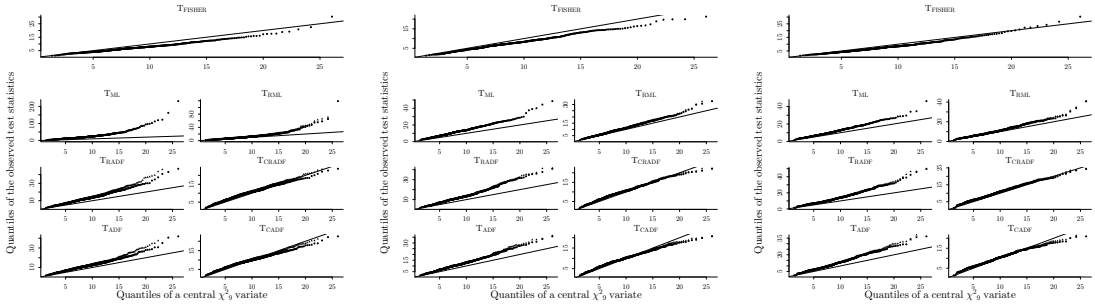
Figure A.2: Q-Q plots $df = 9$ $ICC = .26$



(a) $J = 200$ $n = 10$ (b) $J = 200$ $n = 30$ (c) $J = 200$ $n = 50$



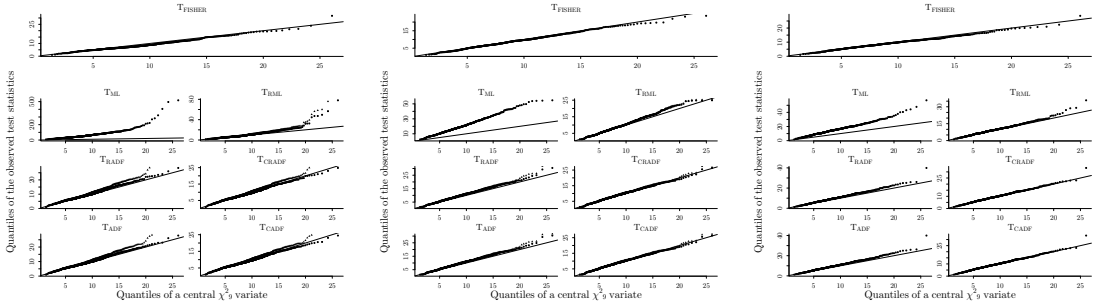
(d) $J = 100$ $n = 10$ (e) $J = 100$ $n = 30$ (f) $J = 100$ $n = 50$



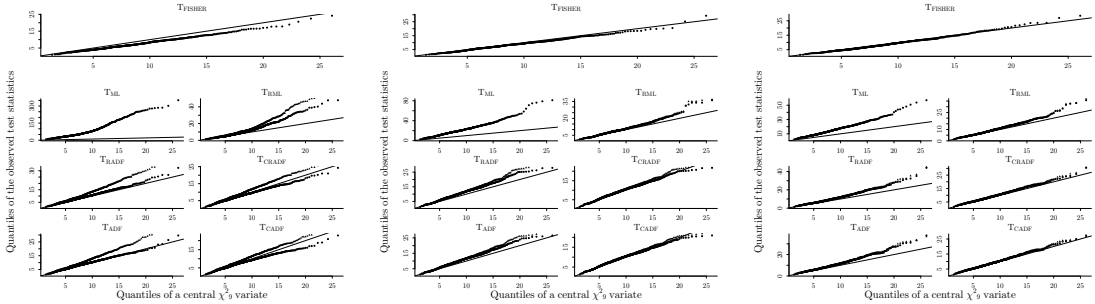
(g) $J = 50$ $n = 10$ (h) $J = 50$ $n = 30$ (i) $J = 50$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

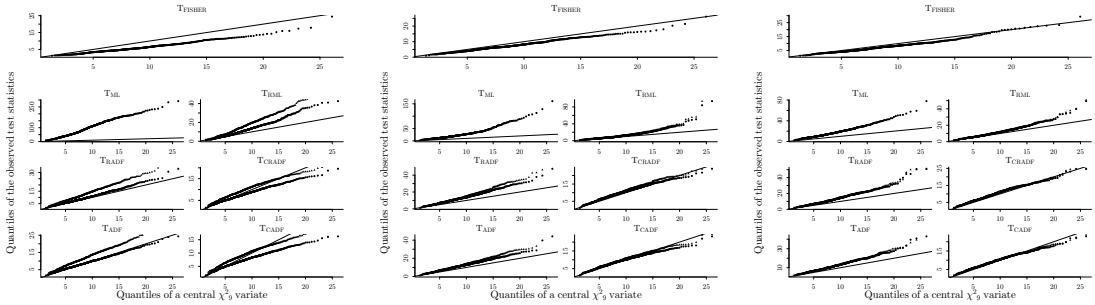
Figure A.3: Q-Q plots $df = 9$ $ICC = .10$



(a) $J = 200$ $n = 10$ (b) $J = 200$ $n = 30$ (c) $J = 200$ $n = 50$



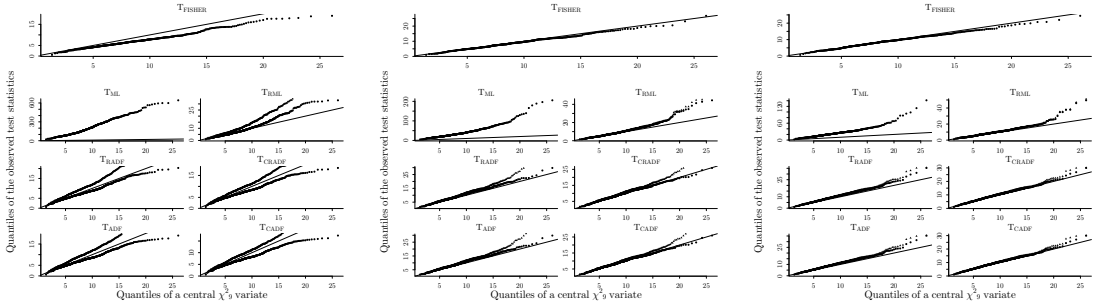
(d) $J = 100$ $n = 10$ (e) $J = 100$ $n = 30$ (f) $J = 100$ $n = 50$



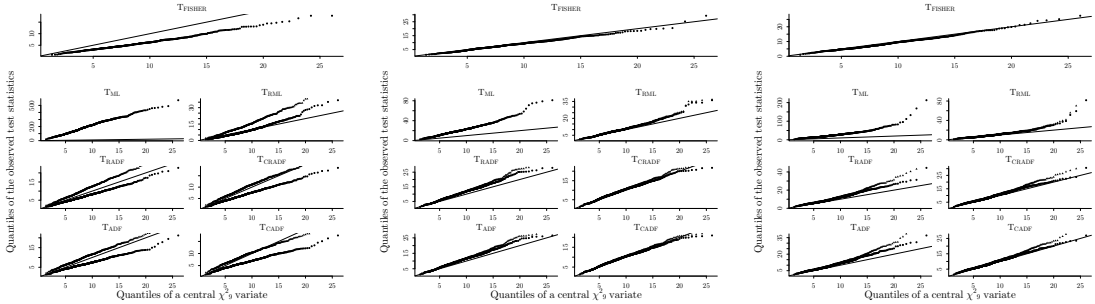
(g) $J = 50$ $n = 10$ (h) $J = 50$ $n = 30$ (i) $J = 50$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

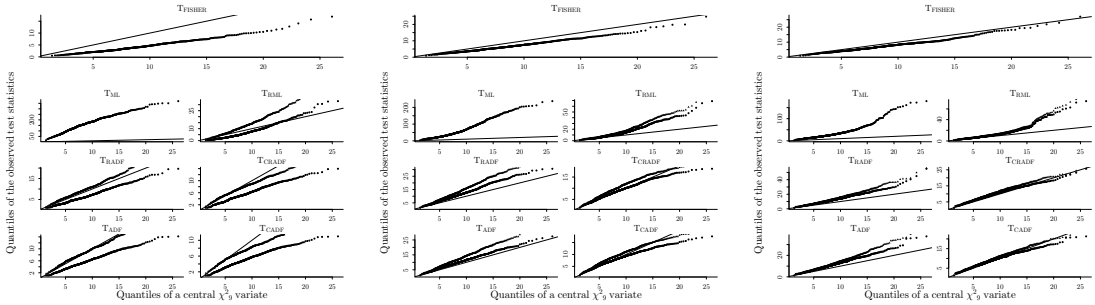
Figure A.4: Q-Q plots $df = 9$ $ICC = .05$



(a) $J = 200$ $n = 10$ (b) $J = 200$ $n = 30$ (c) $J = 200$ $n = 50$



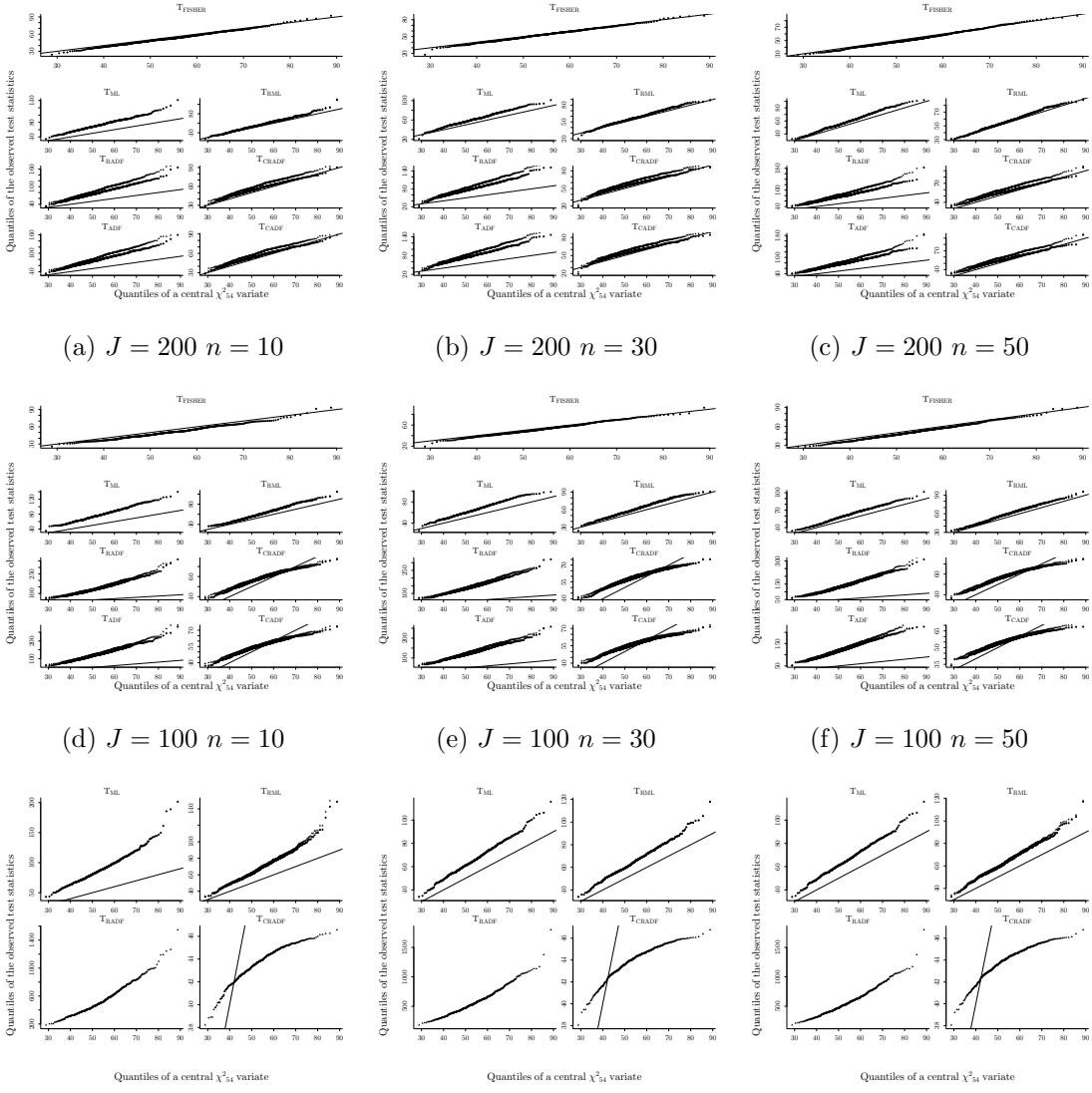
(d) $J = 100$ $n = 10$ (e) $J = 100$ $n = 30$ (f) $J = 100$ $n = 50$



(g) $J = 50$ $n = 10$ (h) $J = 50$ $n = 30$ (i) $J = 50$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

Figure A.5: Q-Q plots $df = 54$, $ICC = .50$



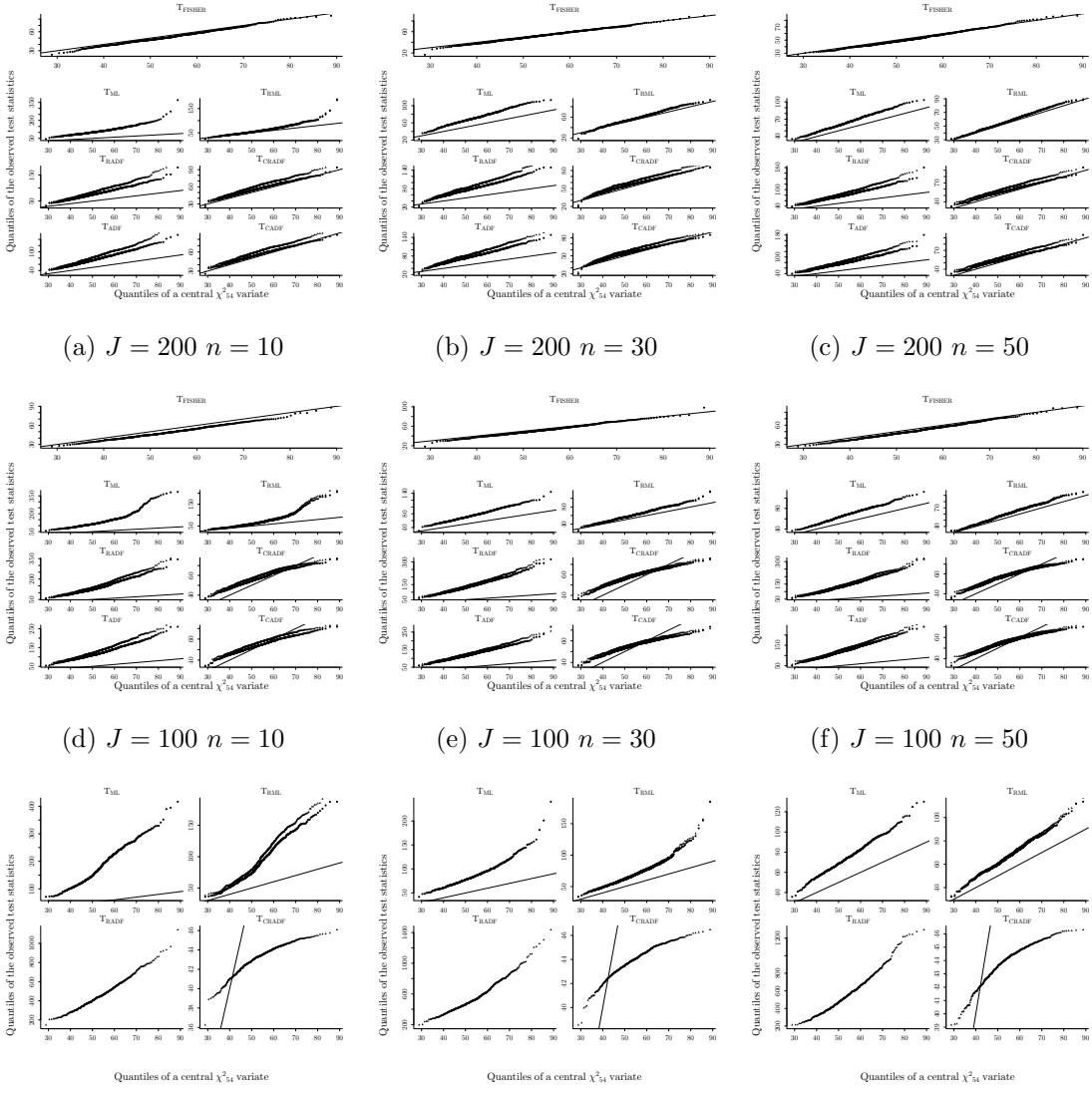
(a) $J = 200$ $n = 10$ (b) $J = 200$ $n = 30$ (c) $J = 200$ $n = 50$

(d) $J = 100$ $n = 10$ (e) $J = 100$ $n = 30$ (f) $J = 100$ $n = 50$

(g) $J = 50$ $n = 10$ (h) $J = 50$ $n = 30$ (i) $J = 50$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

Figure A.6: Q-Q plots $df = 54$ $ICC = .26$



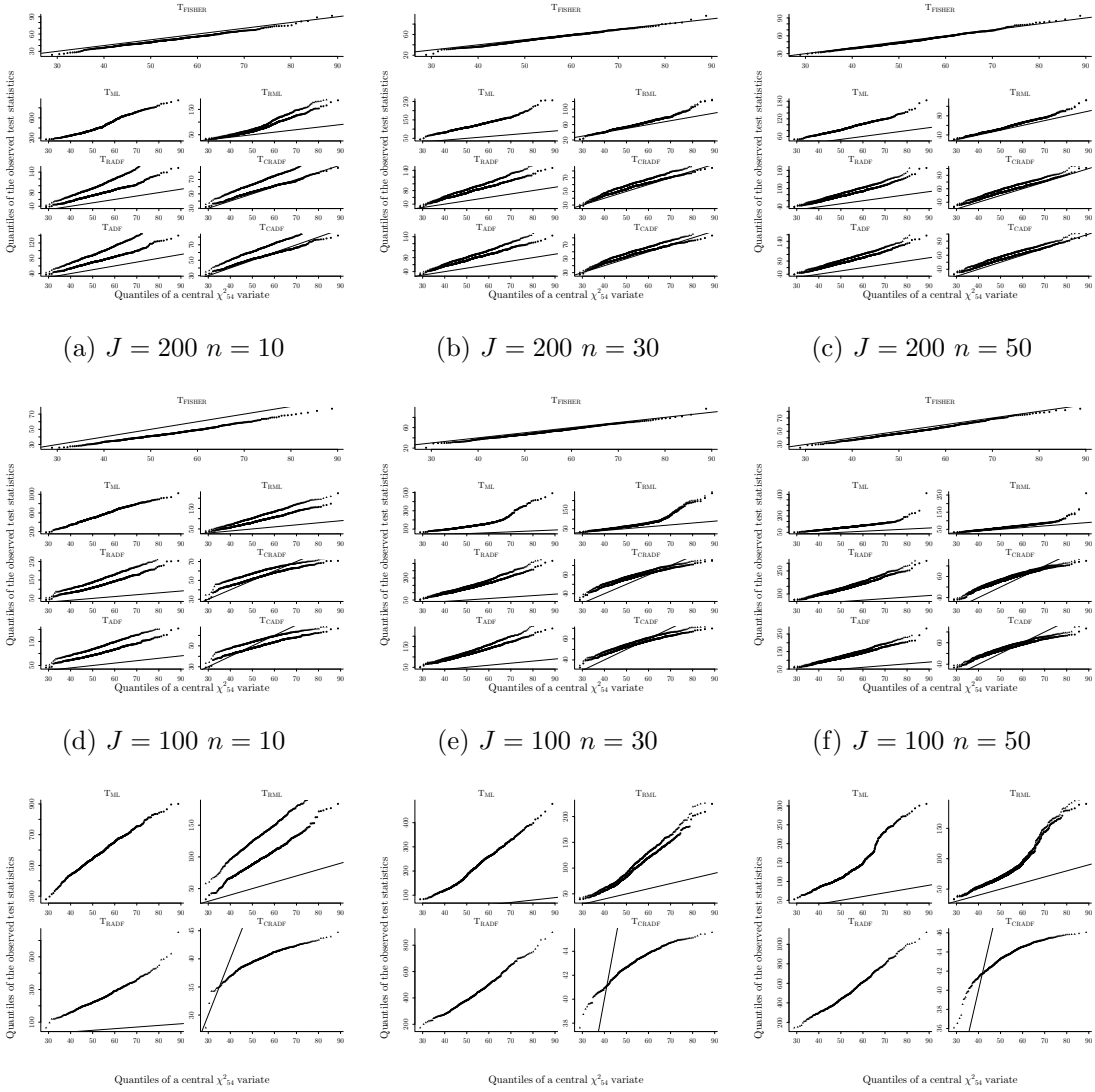
(a) $J = 200$ $n = 10$ (b) $J = 200$ $n = 30$ (c) $J = 200$ $n = 50$

(d) $J = 100$ $n = 10$ (e) $J = 100$ $n = 30$ (f) $J = 100$ $n = 50$

(g) $J = 50$ $n = 10$ (h) $J = 50$ $n = 30$ (i) $J = 50$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

Figure A.7: Q-Q plots $df = 54$ $ICC = .10$



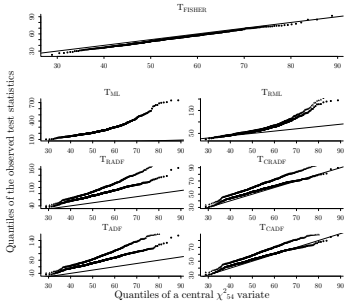
(a) $J = 200$ $n = 10$ (b) $J = 200$ $n = 30$ (c) $J = 200$ $n = 50$

(d) $J = 100$ $n = 10$ (e) $J = 100$ $n = 30$ (f) $J = 100$ $n = 50$

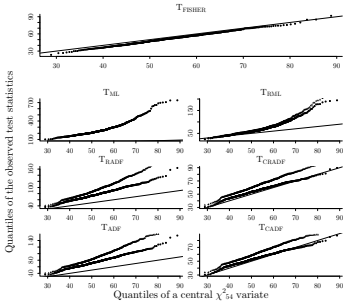
(g) $J = 50$ $n = 10$ (h) $J = 50$ $n = 30$ (i) $J = 50$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

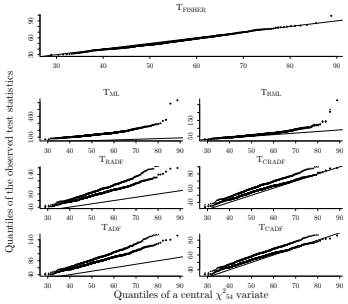
Figure A.8: Q-Q plots $df = 54$ $ICC = .05$



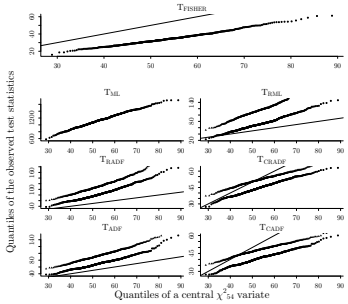
(a) $J = 200$ $n = 10$



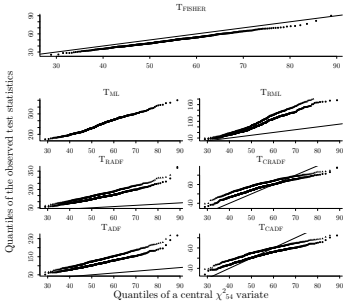
(b) $J = 200$ $n = 30$



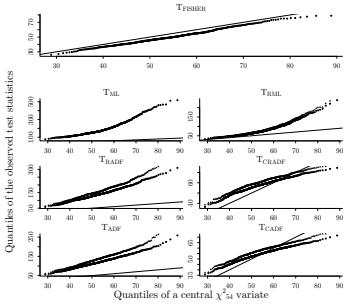
(c) $J = 200$ $n = 50$



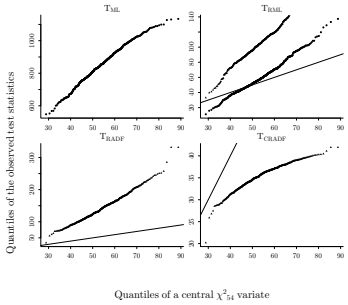
(d) $J = 100$ $n = 10$



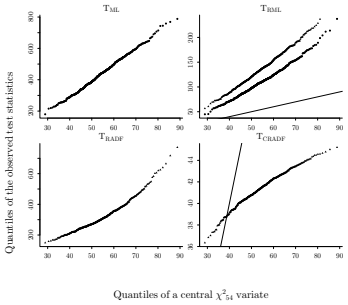
(e) $J = 100$ $n = 30$



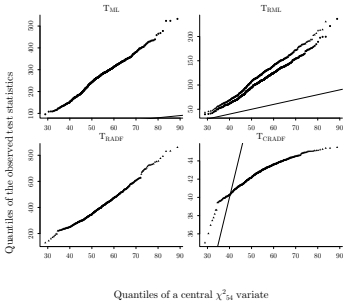
(f) $J = 100$ $n = 50$



(g) $J = 50$ $n = 10$



(h) $J = 50$ $n = 30$



(i) $J = 50$ $n = 50$

Legend: ▲ =Bootstrap, • =GEE

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126.

Akiba, M., LeTendre, G. K., & Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational Researcher*, *36*(7), 369–387.

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning*, *1987*(31), 25–31.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153–166.

Alker, H. R. (1969). A typology of ecological fallacies. *Quantitative Ecological Analysis in the Social Sciences*, 69–86.

American Federation of Teachers. (nd). *A guide for developing multiple measures for teacher development and evaluation.*

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, *37*(2), 65–75.

APA, NCME, AERA. (1999). *Standards for educational and psychological testing.*

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. epi briefing paper# 278. *Economic Policy Institute*.

Balch, R. T. (2012). *The validation of a student survey on teacher practice.* Unpublished doctoral dissertation, Vanderbilt University.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.

Bentler, P. M. (2006). EQS 6 structural equations program manual [Computer software manual]. Los Angeles: BMDP Statistic Software.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*(1), 78–117.

Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. *Multivariate analysis VI*, 9–42.

Bentler, P. M., & Dudgeon, P. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology*, *47*(1), 563–592.

Bentler, P. M., & Liang, J. (2003). Simultaneous mean and covariance structure analysis for two level structural equation models on EQS. In *New developments in psychometrics* (pp. 123–132). Springer.

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*(2), 181–197.

Beran, R., & Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *The Annals of Statistics*, 95–115.

Bill and Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project* (Tech. Rep.).

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis.

Blumberg, A. (1974). *Supervisors and teachers: A private cold war*. McCutchan Publishing Corporation.

Bock, R. D. (1960). Components of variance analysis as a structural and discriminal analysis for psychological tests. *British Journal of Statistical Psychology*, *13*(2), 151–163.

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, *21*(2), 205–229.

Brandt, R. M. (1995). Teacher evaluation for career ladder and incentive pay

programs. In D. L. Duke (Ed.), *Teacher evaluation policy: From accountability to professional development.* SUNY Press.

Braun, H. (2004). *Value-added modeling: What does due diligence require.* Retrieved from `http://www.ncaase.com/docs/Braun2004.pdf`

Braun, H. (2005). Using student progress to evaluate teachers: A primer on value-added models. policy information perspective. *Educational Testing Service*.

Braun, H., Chudowsky, N., Koenig, J., et al. (2010). *Getting value out of value-added: Report of a workshop.* National Academies Press.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, *32*(4), 385–396.

Brennan, R. L. (2001). *Generalizability theory: statistics for social science and public policy.* Springer-Verlag.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 328–375).

Browne, M. (1974). The analysis of patterned correlation matrices by generalized least squares. *British Journal of Mathematical and Statistical Psychology*, *30*(1), 113–124.

Browne, M. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge: Cambridge University Press.

Browne, M. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 1–21.

Browne, M., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403.

Burniske, J., & Meibaum, D. (2011). The use of student perceptual data as a

measure of teaching effectiveness. *Texas Comprehensive Center*.

Bush, R. N. (1954). *The teacher-pupil relationship.* Prentice-Hall.

Camburn, E. M. (2012). Review of "asking students about teaching". *National Education Policy Center*.

Carpenter, J., Goldstein, H., & Rasbash, J. (1999). A non-parametric bootstrap for multilevel models. *Multilevel Modelling Newsletter*, *11*(1), 2–5.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*(2), 234.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189–225.

Cochran-Smith, M. (2010). Forward. In M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook.* John Wiley & Sons.

Colorado Department of Education. (2013). Supporting improved educator evaluations: what is the colorado state model evaluation system.

Connecticut State Department of Education. (2012). Connecticut guidelines for educator evaluation.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis.* Stanford University: Stanford Evaluation Consortium.

Cronbach, L. J., Gleser, G. C., & Nanda, H. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* John Wiley & Sons.

Crosson, A. C., Boston, M., Levison, A., Matsumura, L. C., Resnick, L. B., Wolf, M. K., & Junker, B. W. (2006). Beyond summative evaluation: The

instructional quality assessment as a professional development tool. CSE technical report 691. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

Curran, P. S., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

Darling-Hammond, L. (1990). Teacher evaluation in transition: Emerging roles and evolving methods. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers.* Newbury Park, CA: Sage Publications.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Vol. 1). Cambridge, UK: Cambridge University Press.

Delandshere, G., & Petrosky, A. (2010). The use of portfolios in preservice teacher education: A critical appraisal. *Teacher assessment and the quest for teacher quality: A handbook*, 9–42.

den Brok, P., Stahl, R. J., & Brekelmans, M. (2004). Students' perceptions of teacher control behaviours. learning and instruction. *Psychological Methods*, *14*(4), 425–443.

DePascale, C. A. (2012). Managing multiple measures. *Principal*, *91*(5), 6–10.

Devena, S. E., Gay, C. E., & Watkins, M. W. (2013). Confirmatory factor analysis of the WISC-IV in a hospital referral sample. *Journal of Psychoeducational Assessment*, *31*(6), 591–599.

D'haenens, E., Van Damme, J., & Onghena, P. (2010). Multilevel exploratory factor analysis: illustrating its surplus value in educational effectiveness research. *School Effectiveness and School Improvement*, *21*(2), 209–235.

D'haenens, E., Van Damme, J., & Onghena, P. (2012). Constructing measures for school process variables: the potential of multilevel confirmatory factor analysis. *Quality & Quantity*, *46*(1), 155–188.

DiStefano, C., Monrad, D. M., May, R., McGuiness, P., & Dickenson, T. (2007). Using school climate surveys to categorize schools and examine relationships with school achievement. In *meeting of the American Educational Research Association, Chicago, IL.*

Duncan, A. (2012). *Duncan tells teachers: change is hard.* Retrieved from `http://www.ed.gov/blog/`

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*(1), 149–167.

Eastridge, H. E. (1976). Student evaluation and teacher peformance. *NASSP Bulletin*, *60*(401), 48–54.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Chapman & Hall, New York.

Ewing, J. (2011). Mathematical intimidation: Driven by the data. *Notices of the AMS*, *58*(5), 667–673.

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.

Federal Register. (2009). Part III: Department of Education. , *74*(144).

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, *9*(3), 199–242.

Feng, Z., McLerran, D., & Grizzle, J. (1996). A comparison of statistical methods for clustered data analysis with gaussian error. *Statistics in Medicine*, *15*(16), 1793–1806.

Fenstermacher, G., & Richardson, V. (2005). On making determinations of quality in teaching. *The Teachers College Record*, *107*(1), 186–213.

Ferguson, R. (2010). Student perceptions of teaching effectiveness. *Harvard*

University.

Ferguson, R. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28.

Feuer, M. (2012). No country left behind: Rhetoric and reality of international large-scale assessment. *Princeton, NJ, Educational Testing Service*.

Field, C. A., & Welsh, A. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 369–390.

Flury, B. (1997). *A first course in multivariate statistics*. Springer.

Foldnes, N., Foss, T., & Olsson, U. H. (2012). Residuals and the residual-based statistic for testing goodness of fit of structural equation models. *Journal of Educational and Behavioral Statistics*, *37*(3), 367–386.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal*.

Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2), 271–288.

Gates, B. (2012). *Address at the education commission of the states annual conference.* Retrieved from `gatesfoundation.org/media-center/`

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added* (Tech. Rep.). Mathematica Policy Research.

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review*, *10*(3), 601–616.

Gnanadesikan, R. (1977). Methods for statistical data analysis of multivariate observations.

Goe, L., Holdheide, L., & Miller, T. (2011). A practical guide to designing comprehensive teacher evaluation systems: A tool to assist in the development of teacher evaluation systems. *National Comprehensive Center for Teacher*

Quality.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*(3), 199–208.

Goldstein, H. (2003). *Multilevel statistical models*. New York: John Wiley & Sons.

Good, T. L., & Mulryan, C. (1990). Teacher ratings: A call for teacher control and self-evaluation. *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*, 191–215.

Guion, R. M. (1973). A note on organizational climate. *Organizational behavior and human performance*, *9*(1), 120–125.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis*, *17*(3), 323–336.

Haefele, D. L. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education*, *7*(1), 21–31.

Haertel, E. H. (2013). Reliability and validity idity of inferences about teachers based on student test scores.

Hallquist, M., & Wiley, J. (2013). MplusAutomation: Automating mplus model estimation and interpretation [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=MplusAutomation` (R package version 0.6-2)

Hanushek, E. (2002). Teacher quality. In L. T. Izumi & W. M. Evers (Eds.), *Teacher quality* (Vol. 505). Hoover Institution Press.

Härnqvist, K. (1978). Primary mental abilities at collective and individual levels. *Journal of Educational Psychology*, *70*(5), 706.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? an examination of the statistical properties and policy alternatives. *Education*, *4*(4), 319–350.

Hawaii State Department of Education. (2013). Hawaii educator effective system:

manual for evaluators and participants.

Haycock, K. (2001). Helping all students achieve: Closing the achievement gap. *Educational Leadership*, *58*(6), 6–11.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371–384.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal*, *105*(1), 11–30.

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel. research paper. MET project. *Bill & Melinda Gates Foundation*.

Hoel, P. G., Port, S. C., & Stone, C. J. (1971). *Introduction to statistical theory*. Houghton Mifflin Boston.

Holfve-Sabel, M.-A., & Gustafsson, J.-E. (2005). Attitudes towards school, teacher, and classmates at classroom and individual levels: An application of two-level confirmatory factor analysis. *Scandinavian Journal of Educational Research*, *49*(2), 187–202.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Hoogland, J. J. (1999). *The robustness of estimation methods for covariance structure analysis*. Unpublished doctoral dissertation, University of Groningen.

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

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

Hox, J. J., & Maas, C. (2004). Multilevel structural equation models: The limited

information approach and the multivariate multilevel approach. In *Recent developments on structural equation models* (pp. 135–149). Netherlands: Springer.

Hox, J. J., Maas, C. J., & Brinkhuis, M. J. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*(2), 157–170.

Hoy, W. K., & Clover, S. I. (1986). Elementary school climate: A revision of the ocdq. *Educational Administration Quarterly*, *22*(1), 93–110.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351–362.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 221–233).

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of applied psychology*, *67*(2), 219.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*(1), 85.

John, O. P., & Soto, C. J. (2007). The importance of being valid. *Handbook of research methods in personality psychology*, 461–494.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.

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

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review*

of Educational Research, *47*(2), 267–292.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment. research paper. met project. *Bill & Melinda Gates Foundation*.

Kane, T. J., & Staiger, D. O. (2001). *Improving school accountability measures* (Tech. Rep.). National Bureau of Economic Research.

Kentucky Department of Education. (2012). Kentucky teacher professional growth and effectiveness system: field test guide.

Kozlowski, S. W., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes.

Kratz, H. (1896). Characteristics of the best teacher as recognized by children. *The Pedagogical Seminary*, *3*(3), 413–460.

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction*, *18*(5), 468–482.

Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction*, *40*(2), 44–66.

Ladd, H. F. (2011). Teachers perceptions of their working conditions how predictive of planned and actual teacher movement? *Educational Evaluation and Policy Analysis*, *33*(2), 235–261.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. Yale University Press.

Lawley, D., & Maxwell, A. (1973). Regression and factor analysis. *Biometrika*, *60*(2), 331–338.

Lee, S.-Y. (1990). Multilevel analysis of structural equation models. *Biometrika*,

77(4), 763–772.

Lee, S.-Y., & Poon, W.-Y. (1992). Two-level analysis of covariance structures for unbalanced designs with small level-one samples. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 109–123.

Lee, S.-Y., & Poon, W.-Y. (1998). Analysis of two-level structural equation models via em type algorithms. *Statistica Sinica*, *8*(3), 749–766.

Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology*, *78*(2), 75.

Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika*, *69*(1), 101–122.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lockwood, J., Doran, H., & McCaffrey, D. F. (2003). Using R for estimating longitudinal student achievement models. *R News*, *3*(3), 17–23.

Lockwood, J., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Longford, N., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, *57*(4), 581–597.

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.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, *34*(2), 120–131.

MacCallum, R. C., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative

strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research*, *29*(1), 1–32.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*(3), 502.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. Wiley.

Mair, P., & Wu, E. (2012). REQS: R/EQS interface [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=REQS` (R package version 0.8-12)

Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, *57*(1), 126–134.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253–388.

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.

Martínez, J. F., Borko, H., & Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: lessons learned from two validation studies. *Journal of Research in Science Teaching*, *49*(1), 38–67.

Martínez, J. F., Taut, S., & Schaaf, K. (2013). Classroom observation around the world: Conceptual, methodological, and policy issues for teacher evaluation and development. *manuscript in preparation*.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, *42*(2), 215–232.

Medley, D. M., Coker, H., & Soar, R. S. (1984). *Measurement-based evaluation of teacher performance: An empirical approach*. Longman New York.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156.

Michigan Council for Educator Effectiveness. (2013). Building an improvement-focused system of educator evaluation in michigan: final recommendations.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. (2013). A composite estimator of effective teaching. *Seattle, WA: Bill & Melinda Gates Foundation*.

Millman, J., & Darling-Hammond, L. (Eds.). (1990). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers.* Newbury Park, CA: Sage Publications.

Moir, E. (2009). Validity and reliability of the north carolina teacher working conditions survey. *The University of California at Santa Cruz: New Teacher Center. Retrieved June*, *14*, 2011.

Murphy, J. (1988). Equity as student opportunity to learn. *Theory into Practice*, *27*(2), 145–151.

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

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

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. *Handbook of Advanced Multilevel Analysis*, 15–40.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189.

Muthén, B. O., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 19–30.

Muthén, L. K., & Muthén, B. O. (2010). Mplus: Statistical analysis with latent variables: User's guide [Computer software manual]. Los Angeles: Muthén & Muthén.

National Academies. (2010). *Study of teacher preparation programs in the united states* (Tech. Rep.).

New Mexico Effective Teaching Task Force. (2011). Final report and recommendations.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257.

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods & Research*, *18*(4), 473–504.

Oldham, N. (1974). *Evaluating teachers for professional growth.*

Papay, J. P. (2011). Different tests, different answers the stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163–193.

Partee, G. L. (2012). Using multiple evaluation measures to improve teacher effectiveness: State strategies from round 2 of No Child Left Behind Act

Waivers. *Center for American Progress*.

Patrick, H., Turner, J., Meyer, D., & Midgley, C. (2003). How teachers establish psychological environments during the first days of school: Associations with avoidance in mathematics. *The Teachers College Record*, *105*(8), 1521–1558.

Peterson, K. D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal*, *24*(2), 311–317.

Peterson, K. D. (1995). *Teacher evaluation: A comprehensive guide to new directions and practices*. Corwin Press.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. Corwin Press.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*(2), 135–153.

Phillips, M., & Yamashiro, K. (2013). Reliability and validity of a student survey measuring classroom conditions and practices: Evidence form the pilot administration of the classroom and school environment survey. *paper presented at the AERA annual meeting*.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom assessment scoring system. *Baltimore: Paul H. Brookes*.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and practice*, *16*(2), 9–13.

Popham, W. J. (2013). *Evaluating America's teachers: Mission possible?* Corwin.

Powell, D. A., & Schafer, W. D. (2001). The robustness of the likelihood ratio chi-square test for structural equation models: A meta-analysis. *Journal of Educational and Behavioral Statistics*, *26*(1), 105–132.

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

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Raudenbush, S. W. (2004). Schooling, statistics, and poverty: Can we measure school improvement?. *Educational Testing Service*.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Raudenbush, S. W., Martinez, A., Bloom, H., Zhu, P., & Lin, F. (2010). *Studying the reliability of group-level measures with implications for statistical power: A six-step paradigm* (Tech. Rep.). University of Chicago Working Paper.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the em algorithm and application to us high-school data. *Journal of Educational and Behavioral Statistics*, *16*(4), 295–330.

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics*, *29*(1), 117–120.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, *84*(2), 126–136.

Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X., & Lai, S. (2010). Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, *37*(9), 1487–1498.

Roberts, K. H., Hulin, C. L., & Rousseau, D. M. (1978). Developing an interdisciplinary science of organizations.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 351–357.

Rockoff, J., & Speroni, C. (2008). Reliability, consistency, and validity of the nyc doe environmental surveys: A preliminary analysis. *New York, NY: Department of Education*.

Rosenberg, S. L. (2009). *Multilevel validity: Assessing the validity of school-level inferences from student achievement test data*. Unpublished doctoral dissertation, The University of North Carolina, Chapel Hill.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education*, *4*(4), 537–571.

Rothstein, J., & Mathis, W. J. (2013). Review of "have we identified effective teachers?" and "a composite estimator of effective teaching: Culminating findings from the Measures of Effective Teaching Project". *National Education Policy Center*.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher*, *38*(2), 120–131.

Rowe, K. (2003). The importance of teacher quality as a key determinant of students' experiences and outcomes of schooling. *Building Teacher Quality: What does the research tell us?*, 3.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of educational and behavioral statistics*, *29*, 103–116.

Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents motivation and engagement during middle school. *American Educational Research Journal*, *38*(2), 437–460.

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

Samanta, M., & Welsh, A. (2013). Bootstrapping for highly unbalanced clustered data. *Computational Statistics & Data Analysis*, *59*, 70–81.

Satorra, A., & Bentler, P. M. (1988). *Scaling corrections for chi-square statistics in covariance structure analysis* (Vol. 1).

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction*, *17*(1), 1–24.

Schoenfeld, A. H. (1999). Models of the teaching process. *The Journal of Mathematical Behavior*, *18*(3), 243–261.

Schweig, J. (2013). Cross-level measurement invariance in school and classroom environment surveys implications for policy and practice. *Educational Evaluation and Policy Analysis*. doi: 10.3102/0162373713509880

Schweig, J. (2014). *Multilevel factor analysis by model segregation: comparing the performance of maximum likelihood and robust test statistics*. Unpublished master's thesis, University of California, Los Angeles.

Sexton, J. B., Helmreich, R. L., Neilands, T. B., Rowan, K., Vella, K., Boyden, J., ... Thomas, E. J. (2006). The safety attitudes questionnaire: psychometric properties, benchmarking data, and emerging research. *BMC Health Services Research*, *6*(1), 44.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, *81*(393), 142–149.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24.

Sherman, M., & Cessie, S. l. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized

linear models. *Communications in Statistics-Simulation and Computation*, *26*(3), 901–925.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement*, *17*(4), 245–282.

Song, J., & Felch, J. (2011, 11). *Times updates and expands value-added ratings for los angeles elementary school teachers.*

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems.* RAND Corporation.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors* (Vol. 758). Iowa City, IA.

Stevens, F. I., & Grymes, J. (1993). *Opportunity to learn: Issues of equity for poor and minority students.* US Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics Washington, DC.

Stigler, J., & Stevenson, H. (1992). *The learning gap: Why our schools are failing and what we can learn from japanese and chinese education.* New York: Summit Books.

Tanaka, J. S. (1987). "how big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, 134–146.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *The American Economic Review*, *102*(7), 3628–3651.

The New Teacher Project. (2010). *Teacher evaluation 2.0* (Tech. Rep.).

Thompson, J. E. (1974). Student evaluation of teachers. *NASSP Bulletin*, *58*(384), 25–30.

Timmerman, M. E., Kiers, H. A., Smilde, A. K., Ceulemans, E., & Stouten, J. (2009). Bootstrap confidence intervals in multi-level simultaneous component analysis. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 299–318.

Toland, M. D., & De Ayala, R. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*(2), 272–296.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with" well fitting" models. *Journal of abnormal psychology*, *112*(4), 578.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421–459.

U.S. Department of Education. (2012). *Race to the top—district executive summary.*

Van der Leeden, R., Busing, F., & Meijer, E. (1997). Bootstrap methods for two-level models..

Van Horn, M. L. (2003). Assessing the unit of measurement for school climate through psychometric and outcome analyses of the school climate survey. *Educational and Psychological Measurement*, *63*(6), 1002–1019.

Veldman, D. J., & Peck, R. F. (1969). Influences on pupil evaluations of student teachers. *Journal of Educational Psychology*, *60*(2), 103.

Walker, D. F., & Schaffarzick, J. (1974). Comparing curricula. *Review of Educational Research*, 83–111.

Wang, J. (1998). Opportunity to learn: The impacts and policy implications. *Educational evaluation and policy analysis*, *20*(3), 137–156.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., &

Morgan, K. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360 feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education*, *14*(2), 179–192.

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Prufrock Journal*, *12*(4), 236–247.

Wu, J.-Y., & Kwok, O.-m. (2012). Using sem to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 16–35.

Wu, K. B., Goldschmidt, P., Boscardin, C. K., & Sankar, D. (2009). International benchmarking and determinants of mathematics achievement in two indian states. *Education Economics*, *17*(3), 395–411.

Yang, J. S., & Cai, L. (2012). *Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis-Hastings Robbins-Monro algorithm.*

Yang, J. S., Monroe, S., & Cai, L. (2012). *A multiple group multilevel item bifactor analysis model.*

Yau, L. H.-Y., Lee, S.-Y., & Poon, W.-Y. (1993). Covariance structure analysis with three-level data. *Computational Statistics & Data Analysis*, *15*(2), 159–178.

Yuan, K.-H., & Bentler, P. M. (1997a). Generating multivariate distributions with specified marginal skewness and kurtosis. In W. Bandilla & F. Faulbaum

(Eds.), *Softstat 97-advances in statistical software 6* (pp. 385–391). Lucius and Lucius.

Yuan, K.-H., & Bentler, P. M. (1997b). Improving parameter tests in covariance structure analysis. *Computational Statistics & Data Analysis*, *26*(2), 177–198.

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 289–309.

Yuan, K.-H., & Bentler, P. M. (2002). On normal theory based inference for multilevel models with distributional violations. *Psychometrika*, *67*(4), 539–561.

Yuan, K.-H., & Bentler, P. M. (2003). Eight test statistics for multilevel structural equation models. *Computational Statistics & Data Analysis*, *44*(1), 89–107.

Yuan, K.-H., & Bentler, P. M. (2006). Asymptotic robustness of standard errors in multilevel structural equation models. *Journal of Multivariate Analysis*, *97*(5), 1121–1141.

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*(1), 53–82.

Yuan, K.-H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 397–417.

Yung, Y.-F., & Bentler, P. M. (1994). Bootstrap-corrected adf test statistics in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, *47*(1), 63–84.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, *12*(2), 127–140.