Title
Improving explanatory inferences from assessments

Permalink
https://escholarship.org/uc/item/4fc64449

Author
Diakow, Ronli Phyllis

Publication Date
2013

Peer reviewed|Thesis/dissertation
Improving explanatory inferences from assessments

by

Ronli Phyllis Diakow

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Education in the Graduate Division of the University of California, Berkeley

Committee in charge:
Professor Sophia Rabe-Hesketh, Chair
Professor Mark Wilson
Professor Geoffrey Saxe
Associate Professor Alan Hubbard

Spring 2013
Improving explanatory inferences from assessments

Copyright 2013
by
Ronli Phyllis Diakow
Abstract

Improving explanatory inferences from assessments

by

Ronli Phyllis Diakow

Doctor of Philosophy in Education

University of California, Berkeley

Professor Sophia Rabe-Hesketh, Chair

This dissertation comprises three papers that propose, discuss, and illustrate models to make improved inferences about research questions regarding student achievement in education. Addressing the types of questions common in educational research today requires three different “extensions” to traditional educational assessment: (1) explanatory, trying to explain (or “diagnose”) results regarding students or test items using features of the students and items themselves; (2) longitudinal, modeling change using responses from multiple assessments over time; and (3) multilevel, accounting for higher-level groupings such as classrooms or schools. The papers in this dissertation lie at the intersection of these three areas. Each paper develops a specific statistical or psychometric method with application to educational research.

The first paper proposes and assesses a method for (secondary) data analysis when the outcome variable in a multilevel model is latent and therefore measured with error. The goal is a method that is convenient and readily understandable for applied research. The best current approach for this type of analysis is plausible values methodology, which relies on having a latent regression model for the construction of the plausible values that matches the intended secondary analysis. In current practice, plausible values are constructed with a single-level regression as the conditioning model, which leads to biased estimates of the variance components when the secondary analysis uses a multilevel model. The method proposed in this paper uses weighted likelihood estimates (WLEs) of the latent variable, which do not rely on the specification of a conditioning model, as the dependent variable for the multilevel model. It explicitly accounts for measurement error in the WLEs by fixing part of the level-1 residual variance equal to the estimated variance of the WLEs. The performance of the proposed method is evaluated and compared to the plausible values method using simulation studies and an empirical example.

The second paper proposes extensions to existing item response models for the purpose of evaluating educational interventions. The proposed models incorporate information about the design of the intervention in order to obtain more nuanced information regarding the efficacy of an intervention from the assessment data. The models combine longitudinal
growth on the person side, which provides information about overall efficacy, with group-
and time-varying item feature effects, which provide information about factors that may
contribute to differences in growth over time. The proposed models are applied to empirical
data from a new lesson sequence for elementary school mathematics. Particular attention is
paid to issues of interpretation, item feature design and quality, and measurement invariance.

The third paper proposes a longitudinal item response model for differential growth based
on initial status. The model was designed to answer research questions regarding for whom
an instructional sequence or educational program is more (or less) effective and whether
the instruction or program is expected to narrow or widen an existing achievement gap.
The proposed model encompasses different conceptions of initial status; these conceptions
can be examined simultaneously to uncover whether growth is predicted by factors common
across the assessments or by factors specific to the assessment at the initial time. The
identification and estimation of the proposed model and equivalent models are discussed;
parameter recovery is assessed via simulation. The use and interpretation is illustrated with
empirical data.
To those who fed me along the way.
# Contents

Contents  

List of Figures  

List of Tables  

1 Introduction  
   1.1 Beyond summative assessment  
   1.2 Motivating empirical examples  
   1.3 The three papers  

2 Measurement error in the dependent variable  
   2.1 Introduction  
      2.1.1 Estimators of the latent variable  
      2.1.2 Using estimators in secondary analysis  
   2.2 A multilevel model for secondary analysis with WLEs and fixed variance  
      2.2.1 Model specification  
      2.2.2 When and why this should work (a digression on WLE quality)  
      2.2.3 Relation to other methods to account for measurement error  
   2.3 Simulation studies  
      2.3.1 Simulation 1: Variance components model  
      2.3.2 Simulation 2: Adding covariates  
   2.4 Empirical study  
      2.4.1 PISA 2006  
      2.4.2 Analysis  
      2.4.3 Results  
   2.5 Discussion  
   2.A Appendix  
      2.A.1 Model syntax for Simulation 1  
      2.A.2 Model syntax for Simulation 2  

3 Longitudinal explanatory item response models
3.1 Introduction ........................................................................................................ 64
  3.1.1 Previous models for explanatory and longitudinal purposes ................... 65
  3.1.2 Purposes of this paper ................................................................................... 66
3.2 Longitudinal explanatory models for evaluation ................................................. 67
  3.2.1 Model specification ....................................................................................... 67
  3.2.2 Identification and interpretation .................................................................... 70
  3.2.3 Relation to other models ............................................................................... 73
3.3 Application: Learning Mathematics through Representations (LMR) .............. 74
  3.3.1 Description of the LMR data ........................................................................ 74
  3.3.2 Research questions and analysis ................................................................... 78
  3.3.3 Results ......................................................................................................... 80
3.4 Discussion ........................................................................................................... 94

4 Differential growth item response model ............................................................... 96
  4.1 Introduction ....................................................................................................... 96
  4.2 Longitudinal growth modeling ......................................................................... 97
    4.2.1 Modeling change over time ......................................................................... 98
    4.2.2 Relationship between initial status and change over time ....................... 102
  4.3 A general item response model for differential growth ..................................... 106
    4.3.1 Model specification ..................................................................................... 106
    4.3.2 Interpretation ............................................................................................... 109
    4.3.3 Identification ............................................................................................... 111
    4.3.4 Equivalent models ....................................................................................... 112
    4.3.5 Estimation .................................................................................................... 116
  4.4 Empirical Application ....................................................................................... 120
    4.4.1 Data ............................................................................................................ 120
    4.4.2 Analysis ...................................................................................................... 121
    4.4.3 Results ........................................................................................................ 122
  4.5 Discussion ......................................................................................................... 125
  4.A Appendix ........................................................................................................ 128
    4.A.1 gllamm syntax ............................................................................................ 128
    4.A.2 Mplus syntax .............................................................................................. 129

References ............................................................................................................... 131
List of Figures

2.1 Conditional bias of the WL estimator ........................................ 20
2.2 Conditional asymptotic and empirical variance of the WL estimator ... 21
2.3 Estimated variance components across conditions .......................... 28
2.4 Comparison of bias between methods .......................................... 29
2.5 Interaction diagrams from the ANOVA for the deviations of the person-level variance component from the generating value ....................... 30
2.6 Interaction diagrams from the ANOVA for the deviations of the cluster-level variance component from the generating value .................. 33
2.7 Interaction diagrams from the ANOVA for the deviations of the cluster-level variance component from the simulated abilities estimate .... 34
2.8 Bias across conditions comparing the three PV estimators ................. 39
2.9 Bias across conditions comparing the proposed method and naïve method . 40
2.10 Bias across conditions comparing the proposed method and PVs with dummy variables .......................................................... 41
2.11 RMSE across conditions comparing the proposed method and PVs with dummy variables .......................................................... 42
2.12 Standard errors across conditions comparing the proposed method and PVs with dummy variables .............................................. 43
2.13 Comparison of the secondary analyses to the full measurement model without survey weights ......................................................... 53
2.14 Comparison of the secondary analyses to the full measurement model with survey weights .......................................................... 54
2.15 Standard errors from BRRs versus the Sandwich estimator ............... 56
2.16 Comparison of the different scaling methods for the student-level weights for the full measurement model ........................................ 58
3.1 Example common items from both representation and content domains . 76
3.2 Example non-common items from both representation and content domains . 77
3.3 Results from Model 1: Time-varying, differential item feature effects for common items with effect coding .................................... 84
3.4 Results from Model 3: Time-varying, differential item difficulties for common items with effect coding ........................................ 85
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>Results from Model 4: Time-varying, differential item feature effects for all items with effect coding</td>
<td>87</td>
</tr>
<tr>
<td>3.6</td>
<td>Four example non-common items that are influential for estimating the item feature effects</td>
<td>89</td>
</tr>
<tr>
<td>3.7</td>
<td>Results from Model 7: Time-varying, differential item feature effects for common items with dummy coding (reference items: integers, no line, MC)</td>
<td>91</td>
</tr>
<tr>
<td>3.8</td>
<td>Results from Model 8: Time-varying, differential item feature effects for common items with dummy coding (reference items: integers, line, MC)</td>
<td>92</td>
</tr>
<tr>
<td>4.1</td>
<td>Path diagrams for four latent growth curve models</td>
<td>100</td>
</tr>
<tr>
<td>4.2</td>
<td>Examples of linear growth with different correlations between initial status and rate of change</td>
<td>105</td>
</tr>
<tr>
<td>4.3</td>
<td>Path diagrams for four models for differential growth with different representations of initial status</td>
<td>106</td>
</tr>
<tr>
<td>4.4</td>
<td>Path diagram for the general model for differential growth based on initial status</td>
<td>108</td>
</tr>
<tr>
<td>4.5</td>
<td>Path diagrams for six models that are equivalent</td>
<td>113</td>
</tr>
<tr>
<td>4.6</td>
<td>Path diagram for the general model as implemented in Mplus</td>
<td>117</td>
</tr>
<tr>
<td>4.7</td>
<td>Results for the fixed effects from the simulations for parameter recovery</td>
<td>120</td>
</tr>
<tr>
<td>4.8</td>
<td>Interpretation of the effect of being one or two standard deviations above or below the mean on growth over time</td>
<td>125</td>
</tr>
<tr>
<td>4.9</td>
<td>Example estimated growth trajectories for six illustrative students</td>
<td>126</td>
</tr>
</tbody>
</table>
## List of Tables

2.1 Results from the ANOVA for the deviations of the person-level variance component from the generating value .............................................. 31
2.2 Results from the ANOVA for the deviations of the cluster-level variance component from the generating value .............................................. 32
2.3 Results from the ANOVA for the deviations of the cluster-level variance component from the simulated abilities estimate .................................... 32
2.4 Generating values of the residual variance components and regression coefficients under the relevant simulation conditions ..................................... 36
2.5 Conditioning models for the five different estimators ........................................... 37
2.6 Empirical variables. ..................................................................................... 45
2.7 Comparison of methods, without including survey weights ............................... 52
2.8 Comparison of methods, including fixed final survey weights .......................... 55
2.9 Comparison of different scaling methods for the student-level weights for the full measurement model .................................................. 57
3.1 Description of the eight models fit to the LMR data ................................. 79
3.2 Estimated person regression coefficients (and SEs) from Models 1 to 3 (common items) and 4 to 6 (all items) ........................................... 81
3.3 Estimated item feature effects (and SEs) from Models 1 to 2 (common items) and Models 4 to 5 (all items) .................................................. 82
3.4 Comparison of model fit between the explanatory and descriptive models for the common items .................................................. 86
3.5 Comparison of person regression estimates between effect and dummy coding .................................................................................. 90
3.6 Comparison of item feature effect estimates between effect and dummy coding .................................................................................. 93
4.1 Results from simulations for parameter recovery ............................................. 119
4.2 Estimated growth coefficients and variances (with SEs) from fitting four models to the LMR data .................................................. 123
Acknowledgments

I would never have completed this dissertation without the support and guidance of many people. Though I only mention a small subset of them here, I am very grateful to them all.

I thank Sophia Rabe-Hesketh for her mentorship throughout my years in graduate school and for all her efforts to make my efforts as good as they could possibly be. I thank Mark Wilson for his instruction and counsel, and for always asking the questions with the most salient implications. I thank Geoffrey Saxe for his advice and encouragement, and for providing the opportunity and responsibility for my work to have a meaningful impact on the study of student learning. I thank Alan Hubbard for his critical feedback and for providing an outside perspective.

I thank David Torres Irribarra for his willingness to debate at any time any issue found in psychometrics, education, or on the internet, no matter how trivial or profound, and for his inability to ever let me win an argument. I thank the members of the QME study group for responding to my ridiculous questions by providing insight in addition to laughter. I thank the QME postdocs for their observations about life after graduate school and for helping me get there. I thank Maryl Gearhart, the members of the LMR project, and the RCME fellowship for enabling me to be part of a wider community of education researchers and to place my quantitative research within the larger enterprise.

I thank my parents, Renée Laffer and Nicholas Diakow, for providing me with every opportunity to pursue my dreams, for pushing me to achieve them no matter what they were, and for never saying “I told you so” for categorically refusing to learn statistics in secondary school. I thank my sisters, Marla, Halsey, and Emlyn, for always being there for both celebration and commiseration and for their great attention to detail. And in the end, I thank Douglas London for providing sustenance both emotional and material, for his unending patience, and for never turning down a request for ice cream.
Chapter 1

Introduction

1.1 Beyond summative assessment

Explanatory assessment

Improving the conclusions that can be drawn from assessments underlies the related movements in psychometrics called “explanatory measurement” (De Boeck & Wilson, 2004) and “cognitively diagnostic assessment” (CDA; Nichols, 1994; Nichols, Chipman & Brennan, 1995; Leighton & Gierl, 2007a; Rupp, Templin & Henson, 2010). The full realization of this goal requires coordination between the psychometric models used to generate explanations and inferences and the theory of cognitive development that underlies the domain that is the target of the assessment (Junker, 1999; DiBello, Roussos & Stout, 2007). The ability of assessment to inform and improve instruction and learning relies on the strength of the feedback provided by the assessments (Black & Wiliam, 1998; Pellegrino, Baxter & Glaser, 1999).

Educational assessments must be aligned with theories of cognition in order to provide detailed, targeted information (Nichols, 1994; Pellegrino et al., 1999; Leighton & Gierl, 2007b). It is necessary to have a clearly defined and well-articulated theory of cognitive development both to inform the psychometric modeling and to provide a target of inference in alignment with the purpose of the assessment. Descriptions of the increasingly sophisticated ways that students think about a subject as they learn, called learning progressions, provide an embodiment of a cognitive theory of learning that is well suited to diagnostic or explanatory measurement (National Research Council, 2007; Duncan & Hmelo-Silver, 2009). A similar example is the idea of well-defined, qualitatively-distinct, ordered levels of achievement as the targets of measurement, called construct maps, in the Bear Assessment System (Wilson & Sloane, 2000). Learning progressions and construct maps are formulated to represent development over time and to map links between related aspects in a domain. Both are important to provide useful information for instruction.

The most powerful use of explanatory measurement could be realized when both educational interventions and the related assessments are designed based on a particular theory of
CHAPTER 1. INTRODUCTION

learning. Analyzing these assessments using an appropriate explanatory or diagnostic psychometric model would present the most opportunity for the assessment to provide relevant feedback to the educational process.

Item response theory for assessment

The psychometric models considered in this dissertation belong to the Rasch family of item response models (Rasch, 1960/1980; Adams, Wilson & Wang, 1997), which are a subset of models within item response theory (IRT; Lord, 1980; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; van der Linden & Hambleton, 1997). These models are also located within the more general statistical framework of generalized latent variable models (Bartholomew, 1987; Skrondal & Rabe-Hesketh, 2004, 2007).

Thinking of item response models as generalized latent variable models makes sense for theoretical reasons, since unobservable variables of interest are the targets of assessment, as well as for mathematical reasons, since most item response models, including those in the Rasch family of models, are mathematically equivalent to latent variable models (Mellenbergh, 1994; Adams, Wilson & Wu, 1997; Kamata, 2001; Rijmen, Tuerlinckx, De Boeck & Kuppens, 2003). Placing item response models within the framework of generalized latent variable models allows for very flexible extensions to the models, such as including a variety of person and item covariates (Rijmen et al., 2003; De Boeck & Wilson, 2004). This flexibility enables the models to reflect a detailed and sophisticated understanding of the target of the assessment and to provide better information about what we are assessing.

Good assessment relies on much more than appropriate and sophisticated modeling techniques. Issues such as assessment design, quality, and validity are of primary importance. Though these ideas will not be discussed extensively in this dissertation, the soundness of all of the methods proposed and used here depends on the use of assessment data for which these issues have already been considered.

Multilevel and longitudinal assessment

Assessment data often have a hierarchical structure, because in general students are nested in classrooms, which are nested in schools, which are nested in districts, etc. There are two motivations for accounting for multilevel structure when modeling assessment data: (1) statistical, to account for the dependence induced within the higher-level units, and (2) conceptual, to make inferences about the population of the higher-level units (Raudenbush & Bryk, 2002; Van den Noortgate & Paek, 2004; Rabe-Hesketh & Skrondal, 2012). While the statistical motivation is important for the validity of inferences, herein the conceptual motivation is of more concern.

Educational interventions, particularly those intended for large-scale use, are usually implemented at a higher level of units than individual students. In other words, they are put in place for an entire classroom or school. When evaluating these interventions, we therefore want to be able to make comparisons across the higher-level units (e.g. classrooms
or schools) and to explain those comparisons on the basis of important features either of the
units or of the interventions themselves. In order to do so, the modeling of the assessment
data must incorporate the multilevel structure. This modeling is facilitated by the fact that,
in accord with the previous section, many item response models are in fact multilevel models
(Adams et al., 1997; Kamata, 2001) and can therefore be easily extended to three (or more)
level models (Fox & Glas, 2001; Kamata, 2001; Maier, 2001; Rabe-Hesketh, Skrondal &

Assessment data may also be structured longitudinally. This allows us to examine change
or growth over time. Since educational interventions are designed to produce changes over
time, evaluating the efficacy of these interventions requires modeling change in responses to
assessments over time. Change over time has been modeled via differences in item difficulty
(e.g. within the framework of linear logistic test models; Fischer, 1989; Gluck & Spiel, 1997) or
person ability (e.g. within the framework of multiple dimensions; Andersen, 1985; Embretson,
1991). More recently, models for longitudinal assessment have also been embedded within
a generalized hierarchical model framework (Pastor & Beretvas, 2006), which allows for the
simultaneous modeling of longitudinal variation of items and people. Recent extensions
have also assessed differential growth within latent or manifest groups (Rijmen, De Boeck &

Assessment for educational research

A complete answer to questions common in education research today relies on all three of the
“types” of assessment briefly discussed above: explanatory, multilevel, and longitudinal. The
three papers in this dissertation lie at the intersection of these three areas of item response
modeling. They propose and discuss tools to make more accurate, more specific, and/or
more complete inferences about an educational research question of interest. Though all of
the papers are framed in terms of research in education, they have broader implications as
well.

1.2 Motivating empirical examples

Two empirical data sets provided the impetus for the methods and analysis developed in
this dissertation. The first was a large-scale international assessment that releases the results
from student cognitive assessments for secondary analysis. The second was a research project
relating to the development, implementation, and evaluation of a new curriculum.

Programme for International Student Assessment (PISA)

The Programme for International Student Assessment (PIsA) is a triennial international
program “to assess student performance and to collect data on the student, family and
institutional factors that can help to explain differences in performance” (OECD, 2009b).
The third implementation of PISA was conducted in 2006 (following implementation in 2000
CHAPTER 1. INTRODUCTION

and 2003) and concluded the first cycle of surveys; a second cycle is underway with data collection in 2009, 2012, and 2015.

The PISA assessments focus on fifteen-year-old students’ understanding of reading, mathematics, and science in real-world contexts and on the myriad individual and institutional factors that could influence that understanding. A different cognitive domain (reading, mathematics, and science, respectively) is the primary focus of each implementation, though all of the cognitive domains are assessed on each occasion. PISA also collects information on students’ educational contexts through student and school questionnaires.

The data generated by PISA are publicly available and widely used for secondary data analysis by education researchers. The data that PISA releases include student responses to the cognitive assessment items, plausible values for the students’ performance on the cognitive scales, student responses on the student questionnaire, school responses on the school questionnaire, and indices derived from the responses to the questionnaires. The data also include the sampling strata, survey weights, and a set of replicate weights for the estimation of standard errors.

The empirical data from PISA 2006 used in this dissertation consist of responses to 24 mathematics items from 1177 students in 153 schools in the United States and selected variables for background characteristics of these students and their schools.

Learning Mathematics through Representations (LMR)

The Learning Mathematics through Representations (LMR) project (Saxe, Diakow & Gearhart, 2013) has developed a supplemental curriculum for integers and fractions in the late-elementary grades. The curriculum consists of two two-week sequences of instruction, one for integers and one for fractions. Two central features of the curriculum are (1) the use of the number line as a central representation and (2) the use of a series of principles that reflect the core ideas of the curriculum and are explicitly discussed and agreed upon by the students.

An evaluation of the curriculum was conducted in 2010-2011 (Saxe et al., 2013). Eleven teachers in two San Francisco Bay Area school districts taught the two curriculum sequences in the fall of 2010 (integers in September and October, fractions in November). Students were assessed at four time-points: before the integers lessons (early September), after the integers lessons (mid to late October), after the fractions lessons (early December), and at the end of the academic year (May). The students in ten comparison classrooms where the LMR curriculum was not taught were also assessed at three of the four time-points (early September, early December, May)\(^1\).

The assessments were designed such that about half of the items directly correspond to tasks from the LMR curriculum and half were drawn from other sources including the regular math textbook (which was the same in all classrooms) and released items from large-scale standardized assessments such as the California Standards Tests. Common items were given across the assessments to allow for linking. Dichotomous scoring was used for all items.

\(^1\)Only three assessments were given to the comparison students due to classroom time considerations.
CHAPTER 1. INTRODUCTION

The empirical data from the LMR project used in this dissertation consist of 28-30 item responses\(^2\) for 551 students in 21 classrooms over 4 time points during a single academic year.

1.3 The three papers

This dissertation comprises three papers that propose, discuss, and illustrate models to make improved inferences from assessment data. Each paper develops a specific statistical or psychometric method with application to a question in education research.

In Chapter 2, a method is proposed that accounts for measurement error in the latent dependent variable in multilevel models. The current approach for incorporating a latent variable as the dependent variable in a multilevel model is plausible value methodology (Mislevy, 1991; Mislevy, Beaton, Kaplan & Sheehan, 1992). However, the standard methods for constructing plausible values, such as the one used in PISA, use a conditioning (i.e. latent regression) model that is misspecified with respect to the random-effects structure of the multilevel model, which results in biased parameter estimates (Monseur & Adams, 2009; Diakow, 2010). This paper proposes an alternative method that does not rely on a conditioning model. The proposed method uses weighted likelihood estimates (WLEs; Warm, 1989) of the latent variable as the dependent variable for a multilevel model. It explicitly accounts for measurement error in the WLEs by fixing part of the level-1 residual variance equal to the estimated variance of the WLEs. The approach can be thought of as incorporating a classical measurement model (with non-constant variance) at the lowest level of the multilevel model to account for the estimated, non-constant error variance in the dependent variable. The goal is a method that is both technically accurate and easily implemented for applied research wherein a latent variable is being used as the response variable in a multilevel model.

In Chapter 3, a set of longitudinal explanatory item response models are proposed for the purpose of evaluating educational interventions. These models serve a dual purpose: (1) to account for differential item or facet functioning (DIF/DFF) while estimating an overall longitudinal effect and (2) to use that differential functioning in an explanatory capacity with regards to the overall effect. The proposed models incorporate elements from item response models for longitudinal growth, DIF/DFF, and item parameter drift. They involve extensions to a linear logistic test model (LLTM) that allow the effects of item features to change over time (i.e. includes interactions of item features with time) and also over manifest groups (e.g. treatment conditions) which might be at a higher level such as classroom. Statistically, the proposed models are straightforward extensions; conceptually, they allow for a more fine-grained, explanatory analysis of factors that may contribute to change over time. In particular, when the change is thought to be due to a developmentally-motivated intervention

---

\(^2\)The pretest contained only 28 items; all other assessments contained 30 items, with two additional integers items on the mid-sequence test and two additional fractions items on the posttest. Note that while the tests have common items for linking purposes, they are not identical.
such as LMR, the proposed models allow us to test how item features related to the theory of action of the intervention may contribute to change over time. The paper discusses how the interpretation of time-varying item property effects could be used to make more detailed inferences from well-designed assessment data.

In Chapter 4, a longitudinal item response model is proposed that allows for growth to depend on initial status. These models provide evidence to answer research questions relating to whether students with different levels of prior knowledge (i.e. different initial status) learn at different rates (i.e. exhibit differential growth). The models proposed in this paper extend the HLM model for the relationship between initial status and change (Seltzer, Choi & Thum, 2001, 2003) to include a measurement model for a latent outcome variable; equivalently, the models extend longitudinal growth IRT models (Pastor & Beretvas, 2006; Wilson, Zheng & McGuire, 2012) to include the regression of slope on initial status. The questions that the models in this chapter answer are important for thinking about the consequences of educational programs at all levels. They address for whom the programs are effective and whether the programs are expected to narrow or widen an existing achievement gap.
Chapter 2

Accounting for measurement error in the (latent) dependent variable in multilevel models

2.1 Introduction

In education research, the variables of interest, such as reading achievement or student motivation, often cannot be directly measured or even observed. In order to explore these variables, we use a measurement model, such as an item response model, to provide estimates of the latent variables. If we are interested in the relationship between a latent variable and other variables, the measurement model is only part of what is required for the full analysis.

There are, in general, two ways to use a measurement model to estimate relationships between the latent variable and other variables. The first way incorporates the measurement model at the lowest level of a larger, more complex model that also directly models the desired relationships. Examples of this approach include latent regression item response models (e.g. Mislevy, 1988; Zwindingerman, 1991; Adams et al., 1997; Rijmen et al., 2003) and structural equation models (e.g. Bollen, 1989; Lu & Thomas, 2005). The second way is a two-stage analysis where the primary analysis is the estimation of the measurement model that produces estimates of the latent variable, and the secondary analysis of these estimates either does or does not account for measurement error. The best example of this approach is plausible value methodology (Mislevy, 1991; Mislevy et al., 1992).

Both of these approaches have limitations. Complex models require specialized knowledge and/or software and are often hard to explain to non-methodologists; two-stage plausible value analysis places the burden on the primary data analyst to correctly specify a latent regression model that is general enough for the secondary analysis. When the goal is an ordinary linear regression model with a latent dependent variable, none of these limitations pose an overwhelming challenge, and both approaches are currently widely used in research.

However, when a hierarchical linear model (Raudenbush & Bryk, 2002) with a latent
dependent variable is desired, as has become increasingly commonplace (or even almost mandatory) in education research, the limitations of both approaches are still apparent. Multilevel item response models are computationally demanding and require expert-level knowledge to implement in standard software. The current implementations of plausible values, including large-scale assessments such as the Programme for International Student Assessment (PISA) and the National Assessment of Educational Progress (NAEP), do not use a latent regression model that matches a random-effects multilevel secondary analysis. The use of plausible values obtained using a measurement model with a latent regression that does not account for the multilevel structure can lead to biased estimates for the variance components (Monseur & Adams, 2009; Diakow, 2010); resampling techniques, such as the balanced repeated replications (BRRs) used in PISA, will not obviate this bias. Both PISA and NAEP approximate the random-effects multilevel structure in different ways; PISA includes fixed effects for the higher level units while NAEP includes many covariates at the higher level.

In this paper, we propose an alternative method for incorporating a latent variable as the dependent variable in a multilevel model. The proposed method is a procedure for a secondary analysis that uses weighted likelihood estimates (WLEs; Warm, 1989) of the latent variable as the dependent variable for a multilevel model. WLEs do not require specifying any latent regression model. The proposed method explicitly accounts for measurement error in the WLEs by fixing part of the level-1 residual variances equal to the estimated variances of the WLEs. The approach can be thought of as incorporating a classical measurement model (with non-constant variance) at the lowest level of the multilevel model to account for the estimated, non-constant error variance in the dependent variable. Viewing the proposed method in this way also shows that it is an instance of a level-1 variance-known problem (Raudenbush & Bryk, 2002, chapter 7). Our goal is an accurate method for secondary data analysis that is convenient and readily understandable for applied research.

In the remainder of this section, we describe different estimators for a latent variable and previous results regarding the estimators use in secondary analysis. In Section 2.2, we detail the proposed method and our expectations for its performance. Section 2.3 contains the results from two simulation studies examining the proposed method while Section 2.4 contains an empirical example comparing the proposed method to a multilevel item response model. The paper concludes with a discussion.

2.1.1 Estimators of the latent variable

Consider a latent variable $\theta_p$ that is measured by $I$ dichotomous items with an item response function $P_i(\theta_p)$ that gives the probability of a correct (or affirmative) response for item $i$ given $\theta_p$. Then $Q_i(\theta_p) = 1 - P_i(\theta_p)$ gives the probability of an incorrect (or negative) response.
CHAPTER 2. MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

For example, for the Rasch model (Rasch, 1960/1980), we would have

\[ P_i(\theta_p) \equiv P(Y_{ip} = 1 \mid \theta_p; \delta_i) = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)}, \]  

(2.1)

\[ Q_i(\theta_p) \equiv 1 - P_i(\theta_p) = \frac{1}{1 + \exp(\theta_p - \delta_i)}. \]  

(2.2)

The joint probability of a set of responses \( Y_p \) to the \( I \) items, conditional on the latent variable \( \theta_p \), is given by

\[ L(Y_p \mid \theta_p) = \prod_{i=1}^{I} P_i(\theta_p)^{Y_{ip}} Q_i(\theta_p)^{1-Y_{ip}}. \]  

(2.3)

There are five commonly used methods for obtaining specific values for a latent variable based on the item responses. We use the general term “estimator” to refer to any method of filling in a value (or values) for the latent variable.

**Maximum likelihood (ML) estimate**

If the item difficulties are treated as known (as done in practice), the maximum likelihood (ML) estimate for \( \theta_p \), denoted \( \hat{\theta}_p^{ML} \), is obtained by solving the score equation

\[ \frac{\partial \ln L(Y_p \mid \theta_p)}{\partial \theta_p} = 0. \]  

(2.4)

The asymptotic variance of \( \hat{\theta}_p^{ML} \) is given by

\[ \text{var}(\hat{\theta}_p^{ML} \mid \theta_p) = \frac{1}{I(\hat{\theta}_p^{ML})}, \]  

(2.5)

where \( I(\theta_p) \) is the Fisher information

\[ I(\theta_p) = -E\left( \frac{\partial^2 \ln L(Y_p \mid \theta_p)}{\partial \theta_p^2} \right) = \sum_{i=1}^{I} \frac{[P'_i(\theta_p)]^2}{P_i(\theta_p)Q_i(\theta_p)}. \]  

(2.6)

Lord (1983) showed that \( \hat{\theta}_p^{ML} \) is biased outward given \( \theta_p \) (i.e. the estimates are too large for larger values of \( \theta_p \) and too small for smaller values of \( \theta_p \)).

**Weighted likelihood (WL) estimate**

A weighted likelihood (WL) estimate of \( \theta_p \), denoted \( \hat{\theta}_p^{WL} \), is obtained by solving

\[ \frac{\partial \ln f(\theta_p)}{\partial \theta_p} + \frac{\partial \ln L(Y_p \mid \theta_p)}{\partial \theta_p} = 0, \]  

(2.7)
which corresponds to maximizing a weighted likelihood \( f(\theta_p) L(Y_p | \theta_p) \), where \( f(\theta_p) \) is a penalty function of \( \theta_p \). Warm (1989) selected \( \frac{\partial \ln f(\theta_p)}{\partial \theta} \) such that the bias of \( \hat{\theta}^{\text{ML}}_p \) would be “canceled out”:

\[
\frac{\partial \ln f(\theta_p)}{\partial \theta} = \frac{J(\theta_p)}{2I(\theta_p)},
\]

where \( J(\theta_p) \) is given by

\[
J(\theta_p) = \sum_{i=1}^I \frac{P'_i(\theta_p)P''_i(\theta_p)}{P_i(\theta_p)Q_i(\theta_p)}.
\]

For the Rasch model and the two-parameter logistic model, this equation can be solved, yielding \( f(\theta_p) = \sqrt{I(\theta_p)} \) (Warm, 1989), which corresponds to a Jeffrey’s prior (e.g. Kim & Nicewander, 1993; Magis & Raiche, 2012). For the three-parameter logistic model, there is no closed form solution for \( f(\theta_p) \) (Warm, 1989; Magis & Raiche, 2012).

The asymptotic variance of \( \hat{\theta}^{\text{WL}}_p \) was shown by Warm (1989) to be equivalent to the asymptotic variance of \( \hat{\theta}^{\text{ML}}_p \):

\[
\text{var}(\hat{\theta}^{\text{WL}}_p | \theta_p) = 1 + o(n^{-1}).
\]

Based on a Taylor-series expansion by Warm (1989), Hooijtink & Boomsma (1996) give the asymptotic variance with one more term:

\[
\text{var}(\hat{\theta}^{\text{WL}}_p | \theta_p) = 1 + \frac{J(\hat{\theta}^{\text{WL}}_p)^2}{4I(\hat{\theta}^{\text{WL}}_p)^4} + o(n^{-1}).
\]

Magis & Raiche (2012) give a more accurate estimate as

\[
\text{var}(\hat{\theta}^{\text{WL}}_p | \theta_p) = \frac{1}{I(\hat{\theta}^{\text{WL}}_p) + \frac{J'(\hat{\theta}^{\text{WL}}_p)I(\hat{\theta}^{\text{WL}}_p) + J(\hat{\theta}^{\text{WL}}_p)J'(\hat{\theta}^{\text{WL}}_p) - 2I(\hat{\theta}^{\text{WL}}_p)^2}{2I(\hat{\theta}^{\text{WL}}_p)^2}}.
\]

Most current software that gives \( \text{WL} \) estimates relies on the approximation by Warm (1989). For the Rasch model, the more accurate estimate is straightforward to calculate.

\( \hat{\theta}^{\text{WL}}_p \) was designed and shown by Warm (1989) to be asymptotically unbiased, to order \( o(n^{-1}) \). In finite samples, the conditional bias given \( \theta_p \) is inward (i.e. the estimates are too small for larger values of \( \theta_p \) and too large for smaller values of \( \theta_p \)), a phenomenon also known as shrinkage. The weighting given by Warm (1989) over-corrects for bias in the \( \text{ML} \) estimate.

**Expected/Maximum a posteriori (EAP / MAP) prediction**

An alternative to maximum likelihood estimation of \( \theta_p \) was suggested by Bock & Aitkin (1981) within an empirical Bayesian framework. As for the \( \text{ML} \) and \( \text{WL} \) estimates, the estimates of the parameters are plugged in and treated as known. Predictions for the latent
variable are then based on the estimated posterior distribution of the latent variable given the observed responses, specified by

\[ h(\theta_p \mid Y_p) = \frac{L(Y_p \mid \theta_p)g(\theta_p)}{\int L(Y_p \mid \theta_p)g(\theta_p)d\theta_p}, \tag{2.13} \]

where \( g(\theta_p) \) is the assumed prior density for \( \theta_p \) (e.g. a normal density, as is often assumed). Note that this is not a “new” assumption; the same functional form is usually also assumed during the calibration of the item parameters and the “estimated” prior distribution is then used here. The parameters are set equal to their estimates and treated as known. Bock & Aitkin (1981) suggest two estimates of \( \theta_p \) based on the posterior distribution, one based on the mode and one on the mean.

The maximum a posteriori (map) estimate of \( \theta_p \), denoted \( \hat{\theta}_p^{MAP} \), is obtained by maximizing the estimated posterior distribution in Equation 2.13, that is, from solving:

\[ \frac{\partial \ln h(\theta_p \mid Y_p)}{\partial \theta_p} = 0 \tag{2.14} \]

\[ \frac{\partial \ln g(\theta_p)}{\partial \theta_p} + \frac{\partial \ln L(Y_p \mid \theta_p)}{\partial \theta_p} = 0. \tag{2.15} \]

As seen in Equation 2.15, the map estimate does not require the normalizing constant of \( h(\theta_p \mid Y_p) \). Note that the functions being maximized to calculate the \( \text{WL} \) (Equation 2.7) and \( \text{MAP} \) (Equation 2.15) estimates look similar. However, the form of \( g(\theta_p) \) is very different from \( f(\theta_p) \). We use \( g \) rather than \( f \) here to mark the conceptual distinction between these two estimates (i.e. between the idea of a weighting function and the idea of a prior distribution).

The expected a posteriori (EAP) estimate of \( \theta_p \), denoted \( \hat{\theta}_p^{EAP} \), is obtained by calculating

\[ E[\theta_p \mid Y_p] = \int h(\theta_p \mid Y_p)\theta_p d\theta_p = \frac{\int L(Y_p \mid \theta_p)g(\theta_p)\theta_p d\theta_p}{\int L(Y_p \mid \theta_p)g(\theta_p)d\theta_p}. \tag{2.16} \]

Both the numerator and normalizing constant need to be evaluated numerically. Note that the denominator is the same as the contribution of subject \( p \) to the marginal likelihood.

In finite samples, both the MAP and EAP estimators display shrinkage (i.e. they are biased inward toward the mean and mode of the prior distribution).

**Plausible values (PVs)**

Like the EAP and MAP estimators, plausible values (PVs) are based on the estimated posterior. However, unlike any of the previous methods, PVs do not and are not intended to give point estimates for \( \theta_p \).

As introduced by Mislevy (1991), plausible value methodology addresses the issue of unobserved latent variables by treating them as missing data. Because the latent variable is by definition unobservable, the value of the latent variable is missing for every individual.
The probability of being missing does not depend on any other characteristic of the individual, observed or unobserved, since the probability is 1 for everyone. Therefore, the latent variable is missing completely at random and can be analyzed using the method of multiple imputation for missing data as introduced by Rubin (1987).

The plausible values are the multiple imputations for the latent variable. More formally, they are a set of random draws from the estimated posterior probability distribution of the latent variable for each individual (Mislevy, 1991; Mislevy et al., 1992; Wu, 2005) as given in Equation 2.13. In general, the calculation of the normalizing constant of \( h(\theta_p \mid Y_p) \) is intractable, so numerical methods are used to approximate the posterior distribution and then random draws are taken from the empirical approximation.

For example, consider the method used to draw a set of plausible values for a given person \( p \) by the software ConQuest (Wu, Adams, Wilson & Haldane, 2007), which is used in PISA (OECD, 2009a). Note that although they look similar, this method is not the same as the sampling importance resampling technique (Gelman, Carlin, Stern & Rubin, 2004, pp. 324-325) used to draw plausible values in NAEP (Thomas & Gan, 1997).

First, the estimated posterior \( h(\theta_p \mid Y_p) \) (Equation 2.13) is approximated. A set of \( L \) random values \( \vartheta_l^p \) (PISA uses \( L = 2000 \)) are drawn from \( g(\theta_p) \), the normal prior of \( \theta_p \) with known parameters (in practice, these are estimated in a first stage and then treated as known for the purpose of drawing PVS). These random values are used to calculate the numerator of the posterior

\[
L(Y_p \mid \vartheta_l^p)g(\vartheta_l^p) \equiv f_l^p,
\]

(2.17)

where the other parameters of the likelihood are assumed to be known (in practice, these are estimated in a first stage and then treated as known). The \( f_l^p \) are then normalized by dividing each one by

\[
\hat{s} \equiv \frac{1}{L} \sum_{l=1}^{L} L(Y_p \mid \vartheta_l^m) \approx \int L(Y_p \mid \theta_p)g(\theta_p)d\theta_p.
\]

(2.18)

Note that Equation 2.18 corresponds to Monte-Carlo integration. The posterior is approximated by the pairs \((\vartheta_l^p, \frac{f_l^p}{\hat{s}})\). The probability that \( \vartheta_l^p \) would be drawn is \( p_l^m = \frac{f_l^m}{\sum_{l=1}^{L} f_l^p} \); the cumulative distribution function at \( l \) is approximated by \( \sum_{s=1}^{l} p_s^* \).

Second, a set of \( M \) plausible values is drawn from the approximate density. \( M \) random numbers \( \eta^m \) are drawn from a standard uniform distribution (PISA uses \( M = 5 \)). For each of these random numbers, the \( \vartheta_l^p \) for which the cumulative distribution is closest to \( \eta^m \) without being larger is selected as one of the \( M \) plausible values. This is done by solving the following for \( m_0 \),

\[
\sum_{s=1}^{m_0} p_s^* < \eta^m \leq \sum_{s=1}^{m_0+1} p_s^*.
\]

(2.19)
and taking \( \vartheta_{m0} \) as one of the \( M \) plausible values.

The “missing” latent variable is in this way replaced by a set of plausible values, creating a set of “completed” data sets that can be analyzed using standard procedures for complete data, such as linear regression. Rather than assigning a single estimate for the latent variable, each individual is assigned a vector of plausible values. Note that these PVs are not intended to be the best point estimates, nor even necessarily accurate point estimates, for individuals on the latent variable; nevertheless, they can be used to obtain accurate estimates of population parameters. Most often, five PVs are used for each individual; this number is pervasive though not strongly defended in the literature. While more plausible values would be more accurate, the use of five or even fewer is generally sufficient (Schafer, 1997; Wu, 2005).

In general, the posterior distribution defined by Equation 2.13 is not sufficient for a secondary analysis using PVs. As Mislevy et al. (1992) explains:

We are not estimating population characteristics from plausible values but constructing plausible values to reflect the estimates of population characteristics obtained in a marginal analysis. Unless a characteristic were incorporated into the marginal analysis from which plausible values were constructed, it would generally not be recovered correctly from analyses of the completed data sets.

Thus, PVs are usually drawn from a more complex posterior distribution in which conditioning variables that reflect the variables of interest for the subsequent analysis have been incorporated in the likelihood.

For example, suppose you wanted to conduct a secondary analysis with plausible values to estimate the difference in the mean of the latent variable between males and females. Plausible values constructed as described above assume the same distribution \( g(\vartheta_p) \) for both males and females. Consequently, the posterior distribution used to construct the plausible values will be the same for males and females, the difference in the means of the plausible values will be smaller than the actual difference, and subsequently the secondary analysis will underestimate the difference. The correct difference will be recovered if instead a conditional prior distribution \( g(\vartheta_p \mid \text{gender}) \) is used to define the posterior for constructing the plausible values. While ignoring a covariate when constructing the plausible values will cause bias in the secondary analysis, there is no problem with conditioning on superfluous covariates. The marginal mean of the latent variable is correctly recovered whether the plausible values are constructed conditioning on gender or not.

More generally, if a set of covariates \( X_p \) is desired for the secondary analysis, the model for the latent variable used to construct the plausible values should condition on at least that set of covariates. The model for the latent variable conditional on the covariates is referred to as conditioning model. It usually takes the form of a regression model for \( \vartheta_p \):

\[
\vartheta_p = X_p \beta + \epsilon_p. \tag{2.20}
\]

The use of this conditioning model in conjunction with an item response model, such as the Rasch model (Equation 2.2), is called latent regression item response modeling (Mislevy,
1987; Zwinderman, 1991; Adams et al., 1997). A “correct conditioning model” is defined as a latent regression model having at least all of the the variables that are used in the secondary analysis.

**Comparison of estimators**

A number of studies have compared the point estimators of $\theta$ in terms of their bias and variance (e.g. Warm, 1989; Kim & Nicewander, 1993; Hoijtink & Boomsma, 1995, 1996). PVs are not point estimators, so the comparisons made here do not apply directly; however, since they are drawn from the same posterior distribution that EAP and MAP predictions are obtained from, the results for those estimators apply to the distribution of the PVs.

With a finite number of items, all four of the point estimators are conditionally biased given the true latent variable. The ML estimator is biased outward while the WL, EAP, and MAP estimators are biased inward. In particular, the EAP and MAP are biased towards the mean and mode of the assumed population prior distribution. Comparing the ML, WL, and MAP estimators, Warm (1989) found that the WL estimators were less biased than the others. Similarly, Kim & Nicewander (1993) found that WL estimators were the least biased among the ML, WL, MAP, and EAP estimators. However, both found that this lower bias came at a cost of increased variability, and that the MAP and EAP estimators had a lower variance than the other estimators. The WL estimators also suffered more than the MAP and EAP estimators (through increased bias and variance) when the item parameters (and the variance of $\theta$, though this is not clearly discussed in the original paper) were estimated rather than known (Kim & Nicewander, 1993). All estimators converge to the same limit as the number of items increases and the prior becomes dominated by the likelihood.

**2.1.2 Using estimators in secondary analysis**

Secondary data analysis with any of the point estimators of $\theta_p$ (ML, WL, EAP, MAP) typically proceeds along the lines of data analysis with observed variables.

Secondary data analysis with PVs proceeds along the lines of data analysis for multiply-imputed data. Each completed data set defined by the PVs is analyzed using complete data techniques and then the results are combined into a single result. Following Rubin (1987), Mislevy (1991), and Schafer (1997), the results of an analysis using $M$ previously constructed PVs are combined as follows:

- Obtain an estimate of the parameter $\hat{\beta}_{PV_m}$ and its squared standard error $U_{PV_m}$ by treating each of the $m$ PVs as observed.

- Calculate the estimated parameter $\hat{\beta}_{PV}$ as the average of the estimates of the parameters from each of the $m$ PVs:

$$\hat{\beta}_{PV} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_{PV_m}. \quad (2.21)$$
• Calculate the estimated squared standard error $V_{PV}$ as the sum of the average of the squared standard errors from each of the PV models (within-imputation variance $U_{PV}$) and the variance of the parameters estimates across the different PV models (between-imputation variance $B_{PV}$):

$$V_{PV} = U_{PV} + B_{PV} = \frac{1}{M} \sum_{m=1}^{M} U_{PV_m} + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\beta}_{PV_m} - \hat{\beta}_{PV})^2. \quad (2.22)$$

Population parameters and linear models in secondary analysis

The behavior of the different estimators for a secondary analysis is well-established for estimating population parameters of linear models.

All five of the above-mentioned estimators are unconditionally unbiased for estimating the population mean (if the correct conditioning model is used to construct the posterior underlying the MAP, EAP, and PVs) (Mislevy et al., 1992; Wu, 2005). However, only PVs provide correct estimates of the population variance (Mislevy et al., 1992; Wu, 2005). The variance of the ML and WL estimators is larger than the population variance while the variance of the EAP estimators is smaller. The severity of the bias is worse for shorter tests and for off-target tests (i.e. tests where the item parameters are not well-matched to the distribution of the latent variable).

Similarly, none of the point estimators perform well for examining the relationship between the latent variable and other variables. In an ANOVA analysis, ML, WL, and MAP estimators recover cell means but not cell variances (and therefore not the related sum of squares) (Hoitink & Boomsma, 1996). The ML and WL estimators overestimate the cell variances while the MAP estimators underestimate them. When calculating the correlation between the latent variable and another variable, only PVs with the correct conditioning model recover the generating correlation (Monseur & Adams, 2009). The ML, WL, EAP, and PVs without correct conditioning underestimate the correlation while the EAP estimator with correct conditioning overestimates it. In a secondary linear regression, using EAP estimators or PVs without the correct conditioning model leads to attenuated estimates of regression coefficients (Mislevy, 1991). With the correct conditioning model, the residual variance from EAP estimators will still be wrong, but using PVs leads to estimates similar to those from a full model (Adams et al., 1997).

Thus, PVs with a correct conditioning model represent the best solution for a secondary analysis of a latent variable in linear models. These results are not surprising in light of the fact that PV methodology was specifically designed to ameliorate limitations in using point estimators to estimate population characteristics.

Hierarchical models in secondary analysis

Less work has been done considering the use of estimators of a latent variable in a secondary analysis involving hierarchical linear models, for instance for students nested in schools. We
would expect similar results to those above, with plausible values drawn from a posterior with a correct conditioning model (e.g. that includes random effects for schools) to be the best solution.

Indeed, when Monseur & Adams (2009) conducted simulations to assess the use of different estimators (ML, WL and EAP and PVS with and without conditioning) for recovering variance components in a variance components model, the only estimator that performed well in all conditions and for all parameters of interest were PVS when the conditioning model included the simulated cluster mean of the latent variable. If plausible values were drawn from a measurement model that did not account for the higher level structure, their use in secondary analysis led to biased parameter estimates for the variance components (Monseur & Adams, 2009; Diakow, 2010). However, these simulations provide only an approximation of how PVS could be constructed correctly for real data, as the cluster mean of the latent variable is unknown in practice. Current attempts to approximate the multilevel structure in the conditioning model include the approaches in NAEP, wherein all available covariates at the cluster (school) level are included in the conditioning model, and in PISA, wherein dummy variables for the clusters (schools) are included in the conditioning model.

Verhelst (2010, pp. 211-216) discusses the use of an estimator of the latent dependent variable in secondary analysis. He selects the WL estimator because it is the closest to being conditionally unbiased, which is the requirement for the regression coefficients to be unbiased. From a simulation with a single condition, he finds that when using the WL estimator, based on either generated or estimated item parameters, the parameter estimates are well recovered except for the within-cluster variance, which is overestimated. He does not discuss the impact on the estimated standard errors. Verhelst (2010) warns not to generalize too far from a single condition mimicking a very well-structured assessment. These results suggest that a two-stage analysis is feasible and that the main problem caused by ignoring measurement error is incorrect estimation of the variance components.

A recent study by Li, Oranje & Jiang (2009) proposes an approach for generating plausible values that uses as the conditioning model a hierarchical latent regression model that parallels the multilevel models used for secondary analysis. However, to our knowledge, their methods have not been implemented in practice. In particular, the PVS provided by PISA, NAEP, etc. are not based on multilevel models. Thus, though it is common to use estimators of a latent variable, particularly PVS, as the dependent variable in a multilevel model, this practice is not supported in the literature.

### 2.2 A multilevel model for secondary analysis with WLEs and fixed variance

#### 2.2.1 Model specification

Consider having persons \( p \) within clusters \( c \) for whom the weighted likelihood estimate (WLE) of a latent variable, \( \hat{\theta}_{pc}^{WL} \), and its associated variance, \( \text{var}(\hat{\theta}_{pc}^{WL}) \), have been estimated and
a set of covariates, \( X'_{pc} \), has been observed. The multilevel model for a secondary analysis using the proposed WLE/fixed variance approach can be specified as:

\[
\hat{\theta}^{WL}_{pc} = X'_{pc} \beta + \xi_c + \epsilon_{pc} + \zeta_{pc},
\]

(2.23)

with

\[
\begin{align*}
\xi_c &\sim N(0, \psi^{(2)}) \\
\epsilon_{pc} &\sim N(0, \psi^{(1)}) \\
\zeta_{pc} &\sim N(0, \sigma^2_{pc}) \\
\sigma^2_{pc} &= \text{var}(\hat{\theta}^{WL}_{pc}).
\end{align*}
\]

(2.24) (2.25) (2.26) (2.27)

This model has three error terms: \( \xi_c \), a random intercept at the cluster level (e.g. school or classroom in an education context), and \( \zeta_{pc} \) and \( \epsilon_{pc} \), which are random errors at the person level. This model splits the residual variance (i.e. the variance of the residuals at the person level) into two parts via \( \epsilon_{pc} \), which corresponds to residual error, and \( \zeta_{pc} \), which represents the measurement error. In general, such a model would not be identified because the variance of both \( \zeta_{pc} \) and \( \epsilon_{pc} \) cannot be estimated at the same time. However, we assume that the measurement error variance is known and fix the variance of \( \zeta_{pc} \) to be equal to the estimated variance of the WL estimator (Equation 2.27).

Note that we could rewrite Equation 2.23 as follows:

\[
\hat{\theta}^{WL}_{pc} = \theta_{pc} + \zeta_{pc}
\]

(2.28)

\[
\theta_{pc} = X'_{pc} \beta + \xi_c + \epsilon_{pc}.
\]

(2.29)

This emphasizes the interpretation that the proposed model incorporates a classical measurement model at the lowest level, which is used to account for measurement error, while simultaneously modeling the true value of \( \theta \), \( \theta_{pc} \), as a hierarchical linear model.

WL estimates do not change based on a conditioning model as PVs do. As a result, the proposed method does not rely on having a correctly specified conditioning model. Moreover, the proposed method will not be affected by whether the WL estimates are obtained from a measurement model with no conditioning model, a correctly specified conditioning model, or an incorrectly specified conditioning model.

Estimating the proposed model requires software that can handle (1) three (or more) levels, (2) heteroskedastic variance at the lowest level, and (3) constraints on the variance at the lowest level. Software that meets these requirements, and can therefore be used to estimate the proposed model, include the \texttt{gllamm} command (Rabe-Hesketh, Skrondal & Pickles, 2004b) in Stata (StataCorp, 2011), HLM (Raudenbush, Bryk & Congdon, 2004), and MLwiN (Rasbash, Charlton, Browne, Healy & Cameron, 2009).

\subsection*{2.2.2 When and why this should work (a digression on WLE quality)}

The viability of the proposed method depends on the quality of the WL estimates.
If $\hat{\theta}^{WL}$ is conditionally biased, the parameter estimates from the proposed model will also be biased. In particular, given the inward bias of the WL estimates with a finite number of items, the estimated effects would be attenuated (as seen in the results for linear models discussed above). However, as discussed above, the asymptotic bias of WL estimates is of a lower order than for the other estimators, and simulations have confirmed that WL estimates show less finite-sample bias for a variety of situations. This is the primary reason that WL estimates were chosen for the proposed method rather than the other point estimators.

Between 15 and 25 items have been shown to be sufficient for the bias of the WL estimates to become negligible (Warm, 1989; Hoijtink & Boomsma, 1996; Roberts & Adams, 1997). However, even with 50 items, WL estimates are biased outside the range of the latent variable covered by the item difficulties (Kim & Nicewander, 1993). Thus, in addition to having a sufficient number of items, the WL estimates should also be generated from items whose difficulties have sufficient coverage of the range of the latent variable.

If $\text{var}(\hat{\theta}^{WL} | \theta)$ is incorrect, the proposed model will not appropriately account for the measurement error of $\hat{\theta}^{WL}$. Since the estimation of the variance is based on an asymptotic approximation, it will be incorrect with a finite number of items. Though the WL estimator does not have the minimum variance among the point estimators, this is not required for the method to work well. Rather, the estimated variance must accurately reflect the sampling variance of the estimates.

In general, the asymptotic variance of the WL estimates (Equation 2.10) overestimates the sampling variability of the estimates, with the overestimation being worse for extreme true abilities (Roberts & Adams, 1997). Thus, the proposed method will tend to overcorrect for measurement error. The literature suggests that more than 20-25 items are required for the asymptotic variance to match the sampling variance of the estimates reasonably well (Hoijtink & Boomsma, 1996; Roberts & Adams, 1997). The overestimation will be higher if the approximation in Equation 2.10 is used and less severe if the approximation in Equation 2.12 is used.

Simulations were used to check empirically the quality of the WL estimates and their estimated standard errors from the software to be used for the proposed method. Three factors were manipulated: the number of persons (500, 1000, 5000), the number of items (15, 25, 50), and the range of the item difficulties in relation to the expected person ability distribution (75%, 100%, 125%). 100 replications were run for each of the 27 conditions.

For each person $p$, the person ability $\theta_p$ was drawn from $N(0, 1)$. The item difficulties were generated to be equally spaced across the expected range of the person ability distribution (defined as $[-3, 3]$, wherein 99% of a standard normal distribution is expected to lie) and then multiplied by the appropriate range factor (0.75, 1, or 1.25). For each replication under each condition, item response probabilities were generated for each person to each item following a Rasch model (Equation 2.2) and then item responses were drawn from independent Bernoulli distributions. The Rasch model was fit to the simulated data and the estimated item difficulty parameters were used to calculate the WL estimates and their standard errors as in Equations 2.7 and 2.10.

To consider the quality of the WL estimators, the deviation of the WL estimator was
calculated for each value of $\theta_p$ as $\hat{\theta}^{WL}_p - \theta_p$. Then, the bias $E[(\hat{\theta}^{WL}_p - \theta_p) | \theta_p]$ was examined using loess smoothing and pooling across replications. Figure 2.1 displays plots of the smoothed estimated bias versus the generating ability under each condition. Increasing the number of persons had only a small effect on bias, while the item factors had larger effects. As the number of items and range of item difficulties increases, the conditional bias decreases. With 25 or 50 items whose difficulties cover from -3 to 3 logits, the bias of the WL estimator appears negligible.

To consider the quality of the estimated standard errors, the conditional asymptotic variance of the WL estimator (given by the software using Equation 2.10) was compared to the empirical variance. Pooling across replications, the persons were split into groups of 100 by their generated ability. For each group, the empirical variance of the WL estimator was calculated by taking the variance of the WL estimates while the asymptotic variance was calculated as the square of the average of the reported standard error. Figure 2.2 displays plots comparing the asymptotic and empirical variance for each condition. The number of items appears to be less important for estimating the variance of the WL estimates. With item difficulties covering at least 100% of the person ability distribution, the asymptotic variance appears to approximate the empirical variance well for 15 to 50 items. With only 75% coverage and only 15 items, the asymptotic variance overestimates the empirical variance of the WL estimates. The asymptotic variance should be even closer to the empirical variance if it is calculated following Equation 2.12 rather than Equation 2.10.

These results confirm the findings in the literature regarding the quality of the WL estimates. The number of items and range of item difficulties are important factors when considering the use of the proposed method. We hypothesize that we need at least 25 items with difficulties well-matched to the person ability distribution for the proposed method to work well.
Figure 2.1: Lowess-smoothed conditional bias of the WL estimator as a function of $\theta_p$ under each condition.
Figure 2.2: Conditional asymptotic and empirical variance of the WL estimator as a function of $\theta_p$ for 15 items (top row) and 25 items (bottom row) under each condition.
Figure 2.2: Conditional asymptotic and empirical variance of the WL estimator as a function of $\theta_p$ for 50 items under each condition.
2.2.3 Relation to other methods to account for measurement error

Hoijtink & Boomsma (1996) propose using calculations of the bias and variance of estimates of a latent variable to adjust the results of statistical calculations based on those estimates. In particular, for any statistic based on $\theta$, they instead calculate that statistic for $T - e$, where $T$ is an estimate of $\theta$ and $e$ is the associated standard error. They find that WL estimates work best for this adjustment and demonstrate the adjustments for ANOVA and multiple linear regression. Our method is similar to theirs in considering the WL estimates of the latent variable to arise from a classical measurement model with estimable variance and using that to adjust the secondary analysis. Though the basic idea is similar, our method leverages the hierarchical linear modeling framework to present a method for adjusting for measurement error that is easier to implement.

The proposed method is also similar to the two-stage method proposed by Battauz & Bellio (2011) to account for (latent) covariates measured with error. Battauz & Bellio (2011) use the estimate of the measurement error variance from a measurement model to adjust the desired generalized linear model. They also use WL estimates because they are consistent and asymptotically normally distributed. Though the proposed method incorporates these same features, it differs from the work of Battauz & Bellio (2011) because it is a method for accounting for measurement error in the dependent rather than the independent variable.

The proposed model has a similar mathematical form to common models used for very different purposes, including meta-analysis and small area estimation. The proposed model is essentially a random-effects meta-regression (Raudenbush & Bryk, 1985; Berkey, Hoaglin, Mosteller & Colditz, 1995; Raudenbush & Bryk, 2002, ch. 7) with an additional variance component at the highest level (cf. Thompson, Turner & Warn, 2001). The level-1 model within studies corresponds to the “measurement model” of Equation 2.28 and the level-2 model between studies corresponds to the hierarchical model of Equation 2.29. The proposed model is also mathematically similar to the Fay-Herriot model (Fay & Herriot, 1979) used in small area estimation (Rao, 2003). The goal in small area estimation is to obtain an estimate of a statistic that cannot be directly estimated with precision due to the small population (and therefore sample size). The Fay-Herriot model accomplished this using what they refer to as a indirect estimator that borrows strength from similar units; the model accounts for both between-unit variation (corresponding to $\epsilon_{pc}$ in Equation 2.23) and sampling error with fixed (possibly heteroskedastic) unit-specific sampling variance (corresponding to $\zeta_{pc}$ in Equation 2.23). The proposed method conceptually extends the use of this statistical machinery to a new context in addition to adding a higher level to the models.

The connection to the models for meta-analysis and small area estimation illustrate the connection between the proposed model and a more general framework for considering “second-stage” estimates in the presence of heteroskedasticity (Buonaccorsi, 2006). Applications that Buonaccorsi (2006) considers include random effects and meta-analysis models. Buonaccorsi (2006) concludes that an unbiased estimate of a population mean and its standard error can be obtained by treating the estimators as the unobserved true values. He
derives unbiased estimates of the population variance under various forms of heteroskedasticity in the estimators. These results rely on the assumption of conditionally unbiased estimators with conditionally unbiased standard errors. The use of \( w_l \) estimates in the proposed method reflects these same concerns as discussed in the previous section.

The proposed method matches a previous method used to account for measurement error in the dependent variable by Raudenbush & Sampson (1999). In a paper on incorporating both direct and indirect effects in a multilevel modeling framework, Raudenbush & Sampson (1999) account for measurement error in the outcome variable by including a measurement model at the lowest level in the hierarchical model. As in the proposed method, the latent variable and its measurement error are estimated first in a separate item response analysis (though Raudenbush & Sampson (1999) use a factor analytic approach rather than an item response theory approach). The model is estimated by using the reciprocal of the squared measurement error as a weighting variable at the lowest level and fixing the residual variance at that level to 1. The paper includes no reference to the origin of this method nor discussion of its accuracy.

### 2.3 Simulation studies

Two simulation studies were conducted to examine the performance of the proposed method and compare its performance to that of \( pVs \) with different conditioning models (the current “state of the art”).

When \( pVs \) whose conditioning model does not account for the hierarchical structure are used as the dependent variable in a multilevel model, the within-cluster variance is overestimated while the between-cluster variance is underestimated (Monseur & Adams, 2009; Diakow, 2010). When \( wL \) estimators are used without accounting for measurement error, the within-cluster variance is again overestimated but the between-cluster variance is well recovered (Verhelst, 2010). In both approaches, regression coefficients are well recovered. However, biased estimates of the variance components might affect the estimated standard errors of the regression coefficients.

There were therefore three goals for the simulation studies: (1) to examine the performance of the proposed method with respect primarily to the recovery of the variance components and secondarily to the regression coefficients; (2) to examine the performance of \( pVs \) constructed with conditioning models that account for hierarchical structure in different (though misspecified) ways; and (3) to compare the performance of the proposed method with the performance of the \( pVs \).

We hypothesized that the proposed method using \( wL \) estimators and accounting for measurement error would correct for the bias in the within-cluster variance without biasing the other parameter estimates. We thought that \( pVs \) whose conditioning model included extra covariates at the cluster level would not remove the bias in the variance components because any remaining variability between clusters not explained by the covariates would still be attributed to the individuals. We thought that the \( pVs \) whose conditioning model included
fixed effects for clusters would overcorrect for the observed bias, leading to underestimated within-cluster variance and overestimated between-cluster variance, because all variability between individuals in different clusters is attributed to the cluster.

In the first simulation study, the generating model is a variance components model (i.e. there is no $X_{pc}'\beta$ term in Equation 2.29). The study evaluates the recovery of the variance components under the proposed method. In the second simulation study, the generating model is a random-intercept model with covariates (Equation 2.29). This study evaluates the recovery of the variance components and regression coefficients using different sets of PVs and compares that to the recovery using the proposed method.

In both simulations, the main parameters of interest are the residual variance components. I look at both absolute performance, comparing the results from the proposed method to generating values, and relative performance, comparing the results from the proposed method to (1) an analysis with WL estimates without correcting for measurement error, (2) an analysis with the true generating values, and, in the second simulation (3) an analysis with PVs.

2.3.1 Simulation 1: Variance components model

The purpose of the first simulation study was to examine the performance of the proposed method with respect to the recovery of the variance components. This simulation study used a variance components model (i.e. there is no $X_{pc}'\beta$ term in Equation 2.29) as the generating model.

Conditions

Three main features are hypothesized to impact the usefulness of the proposed method. The first feature is the quality of the WL estimates and associated SE estimates. Two factors were varied to manipulate the quality of the WL estimates: the number of items (15, 25, 50) and the range of the item difficulties in relation to the expected range of the person abilities (75%, 100%, 125%). It was hypothesized that these factors would affect the recovery of both the between- and within-cluster variances, with better recovery when there are more items and a larger item range, because the method should perform better as bias is reduced and the estimated variance approaches the sampling variance.

The second feature is the quality of the variance components estimates. Two factors were varied to manipulate the quality of the estimation of the variance components: the number of clusters (20, 100) and the size of the clusters (25, 50). These factors fixed the number of persons at 500, 1000, 2500, and 5000, respectively. Increases in the number of clusters and the size of the clusters (or possibly their interaction) are hypothesized to affect the recovery of the between-cluster variance, with higher values leading to better estimates.

The third feature is the degree of misspecification of the item response model that does not include a cluster-level random effect. One factor was varied to manipulate the degree of misspecification: the intraclass correlation (0.1, 0.25). The total variance is fixed at
1. We hypothesize that the proposed method will work well regardless of the degree of misspecification.

These five factors resulted in \(3 \times 3 \times 2 \times 2 \times 2 = 72\) conditions under a full factorial design.

**Data generation**

First, item response data were generated. A person ability was generated for each person. With the total person variance fixed at 1, cluster- and person-level random effects were drawn from independent normal distributions with mean 0 and appropriately partitioned variance (given the intraclass correlation) for each person. Then, person ability was calculated as the sum of the random effects. Item difficulties were generated to be equally spaced across \([-3, 3]\) (since the total person variance is fixed at 1, 99% of abilities are expected to lie between -3 and 3 logits) and then multiplied by the appropriate range factor (0.75, 1, or 1.25). Item response probabilities were generated for each person and item following a Rasch model (Equation 2.2) and then item response data were simulated using independent Bernoulli trials.

Second, the WL estimate \(\hat{\theta}_{pc}^{WL}\) and associated standard error \(\text{var}(\hat{\theta}_{pc}^{WL})\) (calculated using Equation 2.12) were obtained after fitting the measurement model

\[
\text{logit}(Y_{ip} = 1) = \theta_{pc} - \delta_i
\] (2.30)

to the item response data.

Third, the proposed model and a naïve model were applied with \(\hat{\theta}_{pc}^{WL}\) as the dependent variable to obtain estimates of the residual variance components. From the proposed model, we obtained \(\hat{\psi}_{WL}\) from

\[
\hat{\theta}_{pc}^{WL} = \beta_0 + \xi_c + \epsilon_{pc} + \zeta_{pc};
\] (2.31)

where

\[
\xi_c \sim N(0, \psi_b) \quad (2.32)
\]
\[
\epsilon_{pc} \sim N(0, \psi_w) \quad (2.33)
\]
\[
\zeta_{pc} \sim N(0, \text{var}(\hat{\theta}_{pc}^{WL})). \quad (2.34)
\]

From the naïve variance components model, which ignored the error in the WL estimate, we obtained \(\hat{\psi}_N\) from

\[
\hat{\theta}_{pc}^{WL} = \beta_0 + \xi_c + \epsilon_{pc};
\] (2.35)

where

\[
\xi_c \sim N(0, \psi_b) \quad (2.36)
\]
\[
\epsilon_{pc} \sim N(0, \psi_w). \quad (2.37)
\]
For comparison, the simulated person abilities were also analyzed as the dependent variable. We obtained $\hat{\psi}_\theta$ from

$$\theta_{pc} = \beta_0 + \xi_c + \epsilon_{pc}, \quad (2.38)$$

where

$$\xi_c \sim N(0, \psi_b) \quad (2.39)$$

$$\epsilon_{pc} \sim N(0, \psi_w) \quad (2.40)$$

$$\zeta_{pc} \sim N(0, \text{var}(\hat{\theta}_{pc}^{WL})). \quad (2.41)$$

There were 20 replications for each of the 72 conditions. Data generation and post simulation analysis was done in R (R Development Core Team, 2012). The item response models were estimated using ConQuest (Wu et al., 2007), and the multilevel models were estimated using Stata (StataCorp, 2011). The proposed model was estimated with gllamm (Rabe-Hesketh et al., 2004b) using adaptive Gauss-Hermite quadrature (Rabe-Hesketh, Skrondal & Pickles, 2005) with eight integration points. The naïve and simulated abilities models were estimated with xtmixed using maximum likelihood estimation. An example of the syntax used to fit each model is given in Appendix 2.A.1.

**Analysis**

Deviations for the variance components under the proposed model were calculated two ways for each replicate $r$: (1) comparing the estimated variance from the proposed model to the generating value ($\hat{\psi}_r^{WL} - \psi$) and (2) comparing the estimated variance from the proposed model to the estimate from the simulated abilities model ($\hat{\psi}_r^{WL} - \hat{\psi}_\theta$), thereby removing some simulation error from the calculation of bias. Calculating the deviations compared to $\hat{\psi}_r^{WL} - \psi$ removes the noise due to simulating $\theta_p$. $\hat{\psi}_\theta$ can be seen as the gold-standard when there is no measurement error; we cannot do better than this with any approach based on item responses. ANOVA was used to compare bias between simulation conditions (Skrondal, 2000). We considered the main effects and two- and three-way interactions of the five factors; $\hat{\psi}_r^{WL} - \psi$ or $\hat{\psi}_r^{WL} - \hat{\psi}_\theta$ was the dependent variable. Effects that were not statistically significant at the 5% level were removed from the model. In addition, the results from the proposed model were compared to results from the naïve model.

**Results**

The estimated variance components from the proposed method are shown in Figure 2.3. Each row of dots displays the estimated variance component from each replication for a single condition. The four dotplots of each color correspond to different numbers of clusters and persons per cluster (from top to bottom: 100 clusters with 50 persons per cluster, 100 clusters with 25 persons, 20 clusters with 50 persons, and 20 clusters with 25 persons). The solid vertical lines are the true values.
Figure 2.3: Estimated variance components across conditions under the proposed method.

The cluster-level variance component is quite well-estimated, though the values are slightly too low. The person-level variance is more severely underestimated. The underestimation becomes smaller as the number of items and range of the item difficulties increases. As the number of clusters and size of clusters increases, the range of the dotplot gets smaller, though the location does not change much. The plots are similar between the two ICC conditions.

Figure 2.4 shows the estimated bias (compared with the generating value) of the variance components estimates for the proposed, naïve, and simulated abilities models. At the person-level, the naïve method overestimates the variance while the proposed method underestimates it, though the magnitude of the bias is worse for the naïve method. At the cluster-level, all three methods recover the generating value quite well with small negative bias, with the naïve method performing slightly better than the proposed method. These results match the findings of Verhelst (2010), who also found that the within-cluster variance is overestimated by the naïve method while the between-cluster variance is recovered.

The results from the ANOVA for the deviations of the person-level (within-cluster) vari-
Bias of Person–level Var.  

Bias of Cluster–level Var.

Figure 2.4: Comparison of bias between the proposed (closed circles), naïve (open circles), and simulated abilities (cross circles) methods.

ance are the same regardless of whether the deviations are calculated with respect to the generating value ($\hat{\psi}_r^{WL} - \psi_r$; Table 2.1) or the results from the simulated abilities model ($\hat{\psi}_r^{WL} - \psi_r^{\theta}$, not shown). The factors related to the items, and therefore to the quality of the WL estimates, are both statistically significant: as the number of items or the range of the item difficulties increases, bias decreases. However, as the number of items increases, the effect of the range of the items decreases (and vice versa; see Figure 2.5a). The effect of the factors related to the number and size of the clusters are borderline significant, but in the expected direction, with more and larger clusters slightly reducing bias (Figure 2.5b). There is less bias for the higher ICC, but as the number of items and the item range increased, the effect of the ICC on the bias becomes negligible (Figure 2.5c).

The results from the ANOVA for the deviations of the cluster-level (between-cluster) variance change depending on whether the deviations are calculated with respect to the gen-
CHAPTER 2. MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

Figure 2.5: Interaction diagrams from the ANOVA for the deviations of the person-level variance component from the generating value, with 95% CIs.
Table 2.1: Results from the ANOVA for the deviations of the person-level variance component from the generating value ($\hat{\psi}_w^{WL} - \psi_w$).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>icc</td>
<td>1</td>
<td>0.08</td>
<td>0.08</td>
<td>31.14</td>
<td>0.0000</td>
</tr>
<tr>
<td>No. Items</td>
<td>2</td>
<td>4.03</td>
<td>2.02</td>
<td>817.02</td>
<td>0.0000</td>
</tr>
<tr>
<td>Item Range</td>
<td>2</td>
<td>0.76</td>
<td>0.38</td>
<td>154.25</td>
<td>0.0000</td>
</tr>
<tr>
<td>No. Clusters</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td>2.99</td>
<td>0.0841</td>
</tr>
<tr>
<td>Cluster Size</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td>3.76</td>
<td>0.0526</td>
</tr>
<tr>
<td>icc:NItem</td>
<td>2</td>
<td>0.02</td>
<td>0.01</td>
<td>4.82</td>
<td>0.0082</td>
</tr>
<tr>
<td>NItem:ItRange</td>
<td>4</td>
<td>0.09</td>
<td>0.02</td>
<td>8.79</td>
<td>0.0000</td>
</tr>
<tr>
<td>NClus:ClusSize</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td>3.19</td>
<td>0.0741</td>
</tr>
<tr>
<td>Residuals</td>
<td>1425</td>
<td>3.52</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

erating value ($\hat{\psi}_r^{WL} - \psi_r$; Table 2.2) or to the results from the simulated abilities model ($\hat{\psi}_r^{WL} - \hat{\psi}_r^{\theta}$; Table 2.3). When bias was calculated by comparison with the generating value, interpretation of the results is unclear. Bias slightly decreases as the number of items increases and as the number of clusters increases. However, there is slightly higher bias for larger icc; in this case, the true variance is larger, so the relative bias may be similar. But, as can be seen in the interaction diagrams (Figure 2.6), no straightforward story emerges. When bias is calculated by comparison with the result from the true model, the effects are more clearly due to the item factors and interactions with icc. Bias decreases as the number of items increases, and the reduction in bias as the number or range of items increases is greater when icc is larger (Figures 2.7a and 2.7b). None of the person factors are significant. Given small numeric values, it appears that using the usual way of calculating bias, simulation error overwhelms or masks effects.

In summary, the proposed method recovers the cluster-level variance quite well, with small negative bias, but underestimates the person-level variance. The bias in the person-level variance is particularly large when the number or range of the items is small, but is less than 0.05 logits with 50 items or an item range of 1.25. The magnitude of the bias for the cluster-level variance is small under all conditions; the estimate is unbiased with 25 or 50 items. Both variance components are recovered well with 25 or 50 items that cover the range of the person abilities well.
Table 2.2: Results from the ANOVA for the deviations of the cluster-level variance component from the generating value ($\hat{\psi}_b^{WL} - \psi_b$).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC</td>
<td>1</td>
<td>0.12</td>
<td>0.12</td>
<td>46.95</td>
<td>0.0000</td>
</tr>
<tr>
<td>No. Items</td>
<td>2</td>
<td>0.06</td>
<td>0.03</td>
<td>11.46</td>
<td>0.0000</td>
</tr>
<tr>
<td>Item Range</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.68</td>
<td>0.5065</td>
</tr>
<tr>
<td>No. Clusters</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td>4.50</td>
<td>0.0341</td>
</tr>
<tr>
<td>Cluster Size</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.8939</td>
</tr>
<tr>
<td>ICC:ItRange</td>
<td>2</td>
<td>0.02</td>
<td>0.01</td>
<td>3.43</td>
<td>0.0325</td>
</tr>
<tr>
<td>NItem:ItRange</td>
<td>4</td>
<td>0.02</td>
<td>0.00</td>
<td>1.86</td>
<td>0.1142</td>
</tr>
<tr>
<td>NItem:ClusSize</td>
<td>2</td>
<td>0.01</td>
<td>0.00</td>
<td>1.19</td>
<td>0.3048</td>
</tr>
<tr>
<td>ItRange:ClusSize</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9972</td>
</tr>
<tr>
<td>NItem:ItRange:ClusSize</td>
<td>4</td>
<td>0.03</td>
<td>0.01</td>
<td>2.64</td>
<td>0.0326</td>
</tr>
<tr>
<td>Residuals</td>
<td>1418</td>
<td>3.48</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Results from the ANOVA for the deviations of the cluster-level variance component from the simulated abilities estimate ($\hat{\psi}_b^{WL} - \hat{\psi}_b^\theta$).

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC</td>
<td>1</td>
<td>0.04</td>
<td>0.04</td>
<td>163.74</td>
<td>0.0000</td>
</tr>
<tr>
<td>No. Items</td>
<td>2</td>
<td>0.04</td>
<td>0.02</td>
<td>103.74</td>
<td>0.0000</td>
</tr>
<tr>
<td>Item Range</td>
<td>2</td>
<td>0.02</td>
<td>0.01</td>
<td>45.28</td>
<td>0.0000</td>
</tr>
<tr>
<td>No. Clusters</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>2.05</td>
<td>0.1524</td>
</tr>
<tr>
<td>Cluster Size</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.12</td>
<td>0.7314</td>
</tr>
<tr>
<td>ICC:NItem</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>10.68</td>
<td>0.0000</td>
</tr>
<tr>
<td>ICC:ItRange</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>9.45</td>
<td>0.0001</td>
</tr>
<tr>
<td>ICC:NClus</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>4.10</td>
<td>0.0430</td>
</tr>
<tr>
<td>Residuals</td>
<td>1427</td>
<td>0.31</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.6: Interaction diagrams from the ANOVA for the deviations of the cluster-level variance component from the generating value, with 95% CIs.
Figure 2.7: Interaction diagrams from the ANOVA for the deviations of the cluster-level variance component from the simulated abilities estimate, with 95% CIs.
2.3.2 Simulation 2: Adding covariates

The purpose of the second simulation study was to compare the performance of the proposed method to the performance of PVS. Because PVS require the specification of a conditioning model and the conditioning models we wished to examine include covariates, this simulation study used a random-intercept model with covariates (Equation 2.29) as the generating model. Performance of each method was evaluated with respect primarily to the recovery of the variance components and secondarily to the recovery of the regression coefficients.

Conditions

The second simulation consisted of a subset of the conditions from the factors of the first simulation (e.g. best and worst). Since the factors related to the items appeared to be the most important, most of the original conditions relating to items were kept: the number of items (15, 25, 50) and the range of the item difficulties in relation to the expected range of the person abilities (75%, 125%). Since the factors related to the persons appeared to mainly affect variability, only the best and worst conditions were kept. A single factor was kept for the number of persons: 20 clusters of size 25 and 100 clusters of size 50 (fixing the number of persons at 500 and 5000, respectively). Both conditions for the fourth factor (intraclass correlation (ICC) of 0.1 and 0.25, with total variance fixed at 1) were carried over from the first simulation, because we hypothesized that the intraclass correlation might matter more for recovery for the PVS-based random intercept models.

An additional factor was manipulated that affects the degree of misspecification: the coefficients of determination due to covariates at both the between- ($\rho^2_b$) and within- ($\rho^2_w$) cluster levels (0, 0.5). The covariates are generated to be orthogonal. $\rho^2_b$ is the proportion of random intercept variance, $\sigma^2_b$, explained by the between-cluster covariate, $x^b_c$, and $\rho^2_w$ is the proportion of the level-1 residual variance, $\sigma^2_w$, explained by the within-cluster covariate, $x^w_{pc}$. Considering the proportional reduction in each of the variance components separately follows Raudenbush & Bryk (2002, chapter 4). Once these factors are set, we can calculate the residual variances ($\psi_b$ and $\psi_w$), residual ICC, and regression coefficients ($\beta_b$ and $\beta_w$) as follows:

- residual variance: $\psi_b = \sigma^2_b(1 - \rho^2_b)$ and $\psi_w = \sigma^2_w(1 - \rho^2_w)$
- residual ICC: $\frac{\sigma^2_b(1 - \rho^2_b)}{\sigma^2_b(1 - \rho^2_b) + \sigma^2_w(1 - \rho^2_w)}$
- regression coefficient: $\beta_b = \rho_b$ and $\beta_w = \rho_w$
- expected SE (Snijders, 2005):
  \[
  \text{var}(\hat{\beta}_b) = \frac{n\psi_b + \psi_w}{mn\text{var}(x^b_c)} \quad \text{and} \quad \text{var}(\hat{\beta}_w) = \frac{\psi_w}{mn\text{var}(x^w_{pc})}\]  
(2.42)
Table 2.4: Generating values of the residual variance components and regression coefficients under the relevant simulation conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Generating Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \rho^2_w )</td>
</tr>
<tr>
<td>ICC</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
</tbody>
</table>

where \( m \) is the number of clusters and \( n \) is the number of persons in each cluster. This factor had three levels; setting both \( \rho^2_b \) and \( \rho^2_w \) to zero was discarded since this would replicate simulation 1. Table 2.4 shows the generating values for each of the simulation conditions related to the ICC and \( \rho^2 \) factors.

These five factors resulted in \( 3 \times 2 \times 2 \times 2 \times 2 \times 3 = 72 \) conditions under a full factorial design.

**Data generation**

The data generation procedure was similar to that from the first simulation.

First, item response data were generated. A person ability was generated for each person. With the total person variance fixed at 1, cluster- and person-level random effects were drawn from independent normal distributions with mean 0 and appropriately partitioned variance (given the intraclass correlation) for each person. Person ability was calculated as the sum of the random effects. Item difficulties were generated to be equally spaced across \([-3, 3]\) (since the total person variance is fixed at 1, 99% of abilities are expected to lie between -3 and 3 logits) and then multiplied by the appropriate range factor (0.75 or 1.25). Item response probabilities were generated for each person to each item following a Rasch model (Equation 2.2) and then item response data were simulated using independent Bernoulli trials.

Second, covariates were calculated by decomposing the generated cluster- and person-level random effects, \( \xi_c \) and \( \epsilon_{pc} \). Candidate covariates \( x^b_c \) and \( x^w_{pc} \) were generated following normal distributions with the same variance as the respective random effects. The final covariates \( x_c^b \) and \( x_{pc}^w \) were then generated by scaling the candidate covariates such that they
Table 2.5: Conditioning models for the five different estimators.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Conditioning model (for ( \theta_{pc} ))</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\theta}_{WL}^{pc} )</td>
<td>( \theta_{pc} )</td>
<td>No conditioning model</td>
</tr>
<tr>
<td>( \hat{\theta}_{PVU}^{pc} )</td>
<td>( \theta_{pc} )</td>
<td>Empty conditioning model</td>
</tr>
<tr>
<td>( \hat{\theta}_{PVM}^{pc} )</td>
<td>( \beta_w x_{pc}^w + \epsilon_{pc}, \beta_b x_{bc}^b + \epsilon_{pc} )</td>
<td>Conditioning model matches analysis model in terms of covariates, ignores multilevel structure</td>
</tr>
<tr>
<td>( \hat{\theta}_{PVC}^{pc} )</td>
<td>( \beta_w x_{pc}^w + \beta_b x_{bc}^b + \epsilon_{pc} )</td>
<td>Conditioning model contains both covariates (regardless of analysis model), ignores multilevel structure (cf. NAEP)</td>
</tr>
<tr>
<td>( \hat{\theta}_{PVD}^{pc} )</td>
<td>( \beta_w x_{pc}^w + x_{pc}^d \beta_d + \epsilon_{pc} )</td>
<td>Conditioning model contains person level covariate and dummy variables for each cluster (cf. PISA)</td>
</tr>
</tbody>
</table>

Third, estimates of the person ability were obtained after fitting a measurement model. In addition to obtaining the \( \hat{\theta}_{WL}^{pc} \) estimate \( \hat{\theta}_{WL}^{pc} \) and associated standard error \( \text{var}(\hat{\theta}_{WL}^{pc}) \) (calculated using Equation 2.12) as in the first simulation, a series of PVs \( \theta_{PV}^{pc} \) with different conditioning models were obtained. The general measurement model fit to the item response data was:

\[
\logit(Y_{ip} = 1) = \theta_{pc} - \delta_i.
\]  

Different conditioning models (different models for \( \theta_{pc} \) in the item response model) were used for the different estimators; these are given in Table 2.5. The WL estimates (\( \hat{\theta}_{WL}^{pc} \)) and naive PVs (\( \hat{\theta}_{PVU}^{pc} \)) used an empty (or null) conditioning model. The conditioning model for the matching PVs (\( \theta_{PVM}^{pc} \)) contained covariates that matched the covariates in the analyzing model. The conditioning model for the PVs with all covariates (\( \theta_{PVC}^{pc} \)) contained both covariates regardless of the analyzing model; this corresponds to the procedure used to generate PVs for NAEP. The conditioning model for the PVs with dummy variables (\( \theta_{PVD}^{pc} \)) contained dummy variables for the clusters and the covariate at the person level; this corresponds to the procedures used to generate PVs for PISA.

Fourth, the proposed model and a series of comparison models with different outcome variables were applied to obtain estimates of the regression coefficients and residual variance components. Using \( \hat{\theta}_{WL}^{pc} \) as the outcome variable, both the proposed model (Equations 2.31
CHAPTER 2. MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

38
to 2.34) and a naïve model (Equations 2.35 to 2.37) were estimated. Using each of the four different sets of PVs, a standard plausible values analysis was performed with results combined as shown in Equations 2.21 and 2.22. Finally, a standard multilevel model was estimated with the true simulated person abilities as the outcome variable. This resulted in a total of seven analysis models. Six estimates were obtained from each of the seven models: three variance components, the person-level, cluster-level, and total residual variances (\( \hat{\psi}_w \), \( \hat{\psi}_b \), and \( \hat{\psi}_w + \hat{\psi}_b \)); and three regression coefficients, the intercept, within-cluster, and between-cluster effects (\( \hat{\beta}_0 \), \( \hat{\beta}_w \), and \( \hat{\beta}_b \)).

There were 20 replications for each of the 72 conditions. The same software was used as in Simulation 1. The proposed model was estimated with \texttt{gllamm} (Rabe-Hesketh et al., 2004b) using adaptive Gauss-Hermite quadrature (Rabe-Hesketh et al., 2005) with eight integration points. The naïve and simulated abilities models were estimated with \texttt{xtmixed} using maximum likelihood estimation. The plausible values models were estimated with \texttt{mi estimate} using \texttt{xtmixed} with maximum likelihood estimation. An example of the syntax used to fit each model is given in Appendix 2.A.2.

Analysis

Deviations from the generating value were calculated for each of the parameter estimates under each of the six analysis models. These deviations were then averaged within each simulation condition to estimate the bias for each parameter. These deviations were squared and then averaged within each simulation condition to estimate the root mean squared error (RMSE) for each parameter. The average of the model-based standard errors was also calculated for the two regression coefficients under each condition.

Results

Figure 2.8 shows the estimated bias under each condition for each parameter comparing three of the PV analyses: with the matched (cross squares), all covariates (open squares), and dummy variables (filled squares) conditioning models. The specification of which condition is given on the left y-axis: at which level a covariate was included, the value of the ICC, the number of items (15 in red, 25 in green, and 50 in blue), and the range of the items (75% is lighter and 125% is darker). The two points within each of these represent the two different numbers of persons and clusters (500 persons in 20 clusters and 5000 persons in 100 clusters). The analysis for the PVs with the null conditioning model is not shown because, as expected, this model did not recover any of the parameters well.

All of the PVs accurately recover the regression coefficients; this is expected since the conditioning models all included the correct covariates. All of the PVs accurately recover the total residual variance; however, they did not accurately partition this variance between the different levels. The PVs that include only the covariates severely overestimate the residual cluster-level variance and underestimate the residual person-level variance. In contrast, the PVs with dummy variables for clusters slightly overestimate the residual cluster-level variance...
and slightly underestimate the residual person-level variance. This effect is mitigated by the increased sample size and is negligible with more than 25 items. PVs where the conditioning model includes dummy variables for clusters in addition to the within covariate perform the best.

These results make sense given the different conditioning models. When the conditioning model includes only covariates, any variability not explained by the covariates is attributed to the individual; as a result, the PVs are drawn from posteriors that do not differ enough in their mean location and are too wide, which decreases their between-cluster variability and increases their between-person within-cluster variability. In contrast, when the conditioning model includes dummy variables for the clusters, all variation between individuals in different clusters is attributed to the cluster; as a result, the PVs are drawn from posteriors that differ too much in their mean location and are too narrow, which increases their between-cluster variability and decreases their between-person within-cluster variability. The magnitude of these effects decreases as the size of the clusters increases. The PVs constructed using dummy
variables for cluster perform much better than the other PVs.

Figure 2.9 shows the estimated bias under each condition for each parameter comparing the proposed method (filled circles) to the naïve method using WL estimates without accounting for measurement error (open circles). The results for the variance components match those from the first simulation.

For the proposed method, the person-level residual variance is underestimated; the magnitude of the underestimation is reduced when the person-level (within) covariate is included. The cluster-level variance is slightly underestimated when the cluster-level (between) covariate is not included, but recovered well when it is. Consequently, the total residual variance is in general underestimated, as in the first simulation. All three of the regression coefficients are recovered well under all conditions, though the within (person-level) covariate is slightly underestimated.

Comparing the proposed method to the naïve method, the proposed method overcorrects for the positive bias in the residual person-level variance under the naïve method, while the
residual cluster-level variance is similarly recovered under both methods. As expected, the naive method provides unbiased estimates of the regression coefficients. Across all conditions, the estimated regression coefficient is lower under the proposed model than the naive model because the correction for measurement error overcompensates for the extra variability in the outcome variable. The proposed method accurately recovers the regression coefficients except for the conditions with few items that do not cover the whole range of the scale, in which cases they are underestimated.

Figure 2.10 displays the estimated bias under each condition for each parameter for the proposed method (circles) and pVs with dummy variables (squares). Comparing these methods, both perform similarly in recovering the cluster-level variance and between regression coefficient. With only 15 items or with items not covering the full range of the scale, the pVs perform better than the proposed method for the person-level variance and the within regression coefficient. Under the other conditions, the two perform similarly for these parameters as well.
Figure 2.11 displays the estimated rmse under each condition for each parameter for the proposed method (circles) and PVs with dummy variables (squares). Both methods have similar variability for recovering the regression coefficients and the cluster-level residual variance. For the person-level residual variance, the two methods are similar with more than 25 items; with only 15 items, the proposed method is much more variable than the PVs with dummies. The smaller sample size increases the variability of the estimates for both methods. In particular, the estimates of the between-cluster regression coefficient and person-level residual variance are very variable with only 500 persons in 20 clusters.

Figure 2.12 displays the average model-based standard errors (sEs) for the two regression coefficients under the proposed method (circles) and PVs with dummy variables (squares). The black vertical lines give the calculated expected SE (Equation 2.42) and the x’s give the model-based standard error calculated using the true simulated values of the latent variable. Note that these are lower bounds to the “true” standard error because measurement error is ignored by considering a model for observed \( \theta_p \). The lower set of standard errors within
within each condition is for the larger sample size; under both models, the $\text{ses}$ are closer to the expected value for the larger sample size. For the within covariate, the proposed method results in larger $\text{ses}$ than the $\text{pv s}$. The difference between them is reduced as the number of items increases; with 50 items, the two methods give similar model-based $\text{ses}$. For the between covariate, the $\text{ses}$ from the $\text{pv s}$ are larger; the difference between them shrinks but is not eliminated as the number of items increases.

In summary, the proposed model works well for estimating both regression coefficients and residual variance components with more than 25 items that cover the full range of the latent variable. The proposed model also gives $\text{ses}$ that are closer to the analysis with the true latent variable under those conditions. With fewer, less well-designed items, the method of plausible values where the conditioning model includes dummy variables at the cluster level appears to be the better approach. Though the $\text{pv s}$ with dummies analysis performs

Figure 2.12: Standard errors across conditions comparing the proposed method (circles) and $\text{pv s}$ with dummy variables (squares).
as well as the proposed method with more items, the proposed method may be preferred because it does not require a correct conditioning model for the covariates. The proposed method can also easily be extended to include random coefficients. It would not be feasible to extend the dummy variable approach to include interactions between dummies and all possible student-level covariates (Li et al., 2009) or to extend the PV methodology to include random coefficients for all possible student-level covariates.

2.4 Empirical study

An empirical analysis was conducted to examine the performance of the proposed method on real data. The primary purpose of the empirical study was to compare the proposed method to the two primary alternatives (using plausible values or using the full multilevel item response model).

2.4.1 PISA 2006

The proposed method is illustrated using the data from the United States sample from the Programme for International Student Assessment (PISA) 2006. PISA is a triennial international program “to assess student performance and to collect data on the student, family and institutional factors that can help to explain differences in performance” (OECD, 2009b). The PISA assessments focus on 15-year-old students’ understanding of reading, mathematics, and science in real-world contexts and on the myriad individual and institutional factors that could influence that understanding. PISA 2006 represents the third of the first cycle of surveys (following PISA 2000 and 2003; a second cycle is underway with PISA 2009, 2012, and 2015), which focused on assessment in reading, mathematics, and science, respectively.

Within the United States, a two-stage stratified sampling design was used to select students (Green, Herget, Rosen & Xie, 2009). At the first stage, schools were selected with probability proportional to size after being stratified based on a number of variables (e.g. public versus private, region of the country, proportion of non-White students). Students were randomly sampled with equal probability at the second stage. The final U.S. sample consisted of 5,611 students in 166 schools in 44 states.

Descriptions of variables

In order to illustrate and discuss the methodology, a set of variables was selected that was thought to be representative of the variables that may be used to answer a substantive question of interest using the methods being studied. The following hypothetical substantive research question was posited in order to select appropriate covariates for the multilevel models:
Controlling for gender, race, and individual-level and school-level socioeconomic status, what is the effect of educational resources at home and at school on student achievement in mathematics?

The variables used in the paper are given in Table 2.6 (OECD, 2009b):

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
<th>Scale or Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome variables:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mresp1 to mresp24</td>
<td>Student responses to the 24 dichotomously scored PISA math items</td>
<td>Dichotomous: 0 Incorrect 1 Correct</td>
</tr>
<tr>
<td>wle</td>
<td>WLE in math constructed by the author from a Rasch model (Equation 2.2)</td>
<td>Logit scale constructed to have mean 0 and standard deviation 1 for the overall latent ability</td>
</tr>
<tr>
<td>pv-naep</td>
<td>PVs in math, five for each student, constructed by the author from a latent regression Rasch model (Equations 2.2 and 2.20) with the six covariates; conditioning model similar to NAEP</td>
<td>Logit scale constructed to have mean 0 and standard deviation 1 for the overall latent ability</td>
</tr>
<tr>
<td>pv-pisa</td>
<td>PVs in math, five for each student, constructed by the author from a latent regression Rasch model (Equations 2.2 and 2.20) with the four student-level covariates and dummy variables for schools; conditioning model similar to PISA</td>
<td>Logit scale constructed to have mean 0 and standard deviation 1 for the overall latent ability</td>
</tr>
<tr>
<td>Independent variables of interest:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hedres</td>
<td>Home educational resources; WL estimates from an IRT scaling using a Rasch model that combines results from seven questions regarding household possessions</td>
<td>Scaled to have mean 0 and standard deviation 1 across OECD countries</td>
</tr>
<tr>
<td>Variable Name</td>
<td>Description</td>
<td>Scale or Categories</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>scmatedu</td>
<td>Quality of school educational resources; WL estimates from an IRT scaling using a Rasch model that combines the results from seven items measuring the school principal’s perceptions of potential factors hindering instruction at school; items were inverted for scaling so more positive values on this index indicate higher levels of educational resources</td>
<td>Scaled to have mean 0 and standard deviation 1 across OECD countries</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>Student self-reported gender</td>
<td>Dichotomous: 0 male 1 female</td>
</tr>
<tr>
<td>race</td>
<td>Student self-reported race</td>
<td>Six categories: 1 Hispanic 2 White, not Hispanic 3 Black, not Hispanic 4 Asian, not Hispanic 5 Multiracial, not Hispanic 6 Other</td>
</tr>
<tr>
<td>escs</td>
<td>Index of economic, social and cultural status PISA 2006; WL estimates from an IRT scaling using a Rasch model that combines results from questions regarding household possessions, number of books in the house, parental occupation, and parental education; used as a measure of student-level socioeconomic status</td>
<td>Scaled to have mean 0 and standard deviation 1 across OECD countries</td>
</tr>
<tr>
<td>perfrlun</td>
<td>Percentage of students eligible for free- or reduced-price lunches in the previous year; used as a measure of school-level socioeconomic status</td>
<td>Range from 0 to 100</td>
</tr>
</tbody>
</table>
The original PISA 2006 U.S. data set contained 5,611 students clustered in 166 schools. The amount of missing data on the covariates considered in this project ranged from 0% to 7.1%. After list-wise deletion using these covariates, the data set contained 5,056 students clustered in 153 schools; the mean number of students per school was 33.0 (standard deviation 5.4). As discussed below, the data were subset further to retain only students who answered booklets 2, 10, or 11. The final data set for this project contained 1177 students clustered in 153 schools; the mean number of students per school was 7.7 (standard deviation 1.8), with a minimum of 1 and a maximum of 11. PISA uses a matrix sampling design for the assessments wherein the items are distributed among thirteen booklets (with each item appearing in four of the booklets) and each student answers only the subset of items found on one of the booklets. 383 students received a booklet with all 24 items and 794 students received booklets with 12 of the items; only 2 students were not scored on all of the items in their booklet.

### Sampling weights

The PISA data contain two types of sampling weights: (1) final survey weights at the school and student levels and (2) replicate weights (balanced repeated replications; BRRs) at the student level. The purpose of the survey weights is to provide valid inference to the target population in light of the unequal (though random) selection of schools and students and differential response rates. The purpose of the replicate weights is to estimate standard errors (rather than using, for example, model-based standard errors). For a detailed description of the calculation of the weights, see OECD (2009a).

The final sampling weights are used to adjust the analysis for inference to the population since students were chosen randomly but with different selection probabilities. When the analysis model is a multilevel model, separate weights are needed at each level to account for the selection probabilities at each stage (Rabe-Hesketh & Skrondal, 2006). PISA provides separate weights at each level. The final school level weights are the school base weight (the reciprocal of the school’s probability of selection) adjusted for school non-response and trimmed\(^1\). The final student level weights are the student base weight (the reciprocal of the student’s probability of selection within their school) adjusted for student non-response and trimmed\(^1\).

If raw (unscaled) weights are used at level 1 (in this example, at the student level), the between-cluster (in this example, between-school) variance is overestimated (Pfefferman, Skinner, Holmes, Goldstein & Rasbash, 1998; Rabe-Hesketh & Skrondal, 2006). Two common scaling methods for the level-1 weights are (1) scaling the weights to sum to the actual cluster size (hereafter called size) and (2) scaling the weights to sum to the effective sample size (hereafter called effective). Both of these scaling methods overcorrect for the bias in the

\(^1\)In the United States sample, only one school weight and no student weight required trimming (Green et al., 2009).
between-cluster variance, leading to potentially underestimated between-cluster variance. In addition, for a generalized linear mixed model (as opposed to a linear mixed model), the estimated regression coefficients are biased when the estimated variance component is biased (Rabe-Hesketh & Skrondal, 2006). Bias is expected for small clusters or informative level-1 weights. The sensitivity of the analysis to the scaling of the weights can be checked by comparing the analysis with different scaling methods.

Though the subset of PISA data considered here has small clusters (average size 7.7), the student level weights are non-informative (the weights rescaled under both methods have mean 1.00 and standard deviation 0.05). The bias was therefore expected to be small. To check, the analysis was run using both the size and effective rescaling methods in addition to the raw weights.

When using survey weights, correct standard errors can be obtained using the Sandwich estimator, sometimes referred to as linearization (White, 1982; Rabe-Hesketh & Skrondal, 2006). Alternatively, the replicate weights are used to calculate the sampling variance in the parameter estimates due to the sampling of schools and students. The use of replicate weights in this manner is similar to the use of resampling procedures such as the bootstrap or jackknife to calculate standard errors. The method of BRRs is most similar to the jackknife, but at each replication half of the observations are removed (rather than just one observation). Since completely ignoring half of the available information at each replication would result in a great loss of efficiency, Fay’s (1984; 1989) variant of BRRs inflates and deflates each half of the observations under each replication rather than removing them entirely.

PISA uses the method of BRRs with Fay’s variant with a factor of 0.5 (for a detailed description of the general method, see pages 139-141 in OECD (2009a) and pages 70-74 in OECD (2009b)). To construct the BRRs, schools are paired based on the sampling strata; this pairing accounts for the stratification of the original sampling design. Within each pair, the sample from one school was inflated and the other deflated by multiplying the school base weight by 1.5 or 0.5 respectively. Thus, resampling is of whole schools, since these were the primary sampling unit. The determination of which schools were inflated or deflated is determined by a Hadamard matrix of order 80, resulting in 80 replicate weights. The non-response and trimming adjustments were then recalculated for each replicate.

Using the replicate weights to calculate the sampling variance for a parameter requires running the analysis using the final weights to obtain an estimate of the parameter $\hat{\beta}$ and then using each replicate weight separately to obtain $\hat{\beta}_r$ for each replication $r$. Then, the sampling variance is calculated as:

$$V_{BRR} = \frac{1}{G(1 - k)^2} \sum_{r=1}^{G} (\hat{\beta}_r - \hat{\beta})^2$$  \hspace{1cm} (2.46)

where $G = 80$ is the number of replicates and $k = 0.5$ is the Fay’s factor.

\footnote{Different weighting factors are used if any of the pairs are instead triples; however, the United States sample contained no triples.}
CHAPTER 2. MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

The PISA 2006 data provided by the OECD contains only student level replicate weights. However, the primary resampling was at the school level. In addition, separate weights are needed at each level for analyses using multilevel models. Thus, replicate weights at the school level were reconstructed as follows. First, the weighting factor for each school for each replication was recovered by dividing each student level replicate weight by the final student level weight and rounding to the nearer of 0.5 or 1.5. Then, the final school level weight was multiplied by the weighting factor for each replication, resulting in replicate school level weights. These reconstructed school level replicate weights were used in the analysis in conjunction with the provided student level replicate weights.

Simplifications

Three simplifications were made for the empirical analyses for this paper. Each of these simplifications represents a complexity in the PISA data that can be ignored for the purposes of illustrating the specific methodology discussed in this paper. While these simplifications prevent this analysis from being used to make inferences regarding the population that the PISA data were intended to represent, they do not bias the methodological comparisons that are the focus of this paper. A more robust analysis that accounts for these complexities would be needed to answer substantive questions using the PISA data.

The first two simplifications are commonly made when analyzing the PISA data. First, missing values in the covariates were dealt with via list-wise deletion. A more robust analysis would consider whether the missing values are missing (completely) at random, and would perhaps use an imputation method to prevent data loss. This simplification was undertaken mainly to limit the number of imputed data sets under consideration to the five corresponding to the plausible values. Second, the scale scores for the indices of educational resources and socioeconomic status were treated as observed continuous variables. A more robust analysis would take into account that each of these indices in fact represents an unobserved latent variable and is therefore subject to measurement error.

The third simplification is less common. The analysis was restricted to the items and students for three of the thirteen test booklets (booklets 2, 10, and 11). This simplification addressed two complexities in the full PISA data set: first, the large number of items, some of which were dichotomous and some of which were partial credit; second, the large amount of missing data due to the matrix sampling design (including many students who did not answer any mathematics items at all). Booklets 2, 10, and 11 were selected because they were the three booklets that contained only dichotomously scored items. The decision was made to use three booklets rather than only one because each booklet was only given to a few (1-5) students per school, which resulted in cluster sizes that were too small at the school level. With the small cluster sizes, the data would have been too sparse for the multilevel logistic regression models; small cluster sizes also reduce power and result in unstable estimates.
2.4.2 Analysis

The purpose of the analysis was to compare the results for real empirical data between the proposed method for secondary analysis using weighted likelihood estimates (WLEs) with a fixed heteroskedastic variance component to two alternatives, (1) a plausible values secondary analysis and (2) a full multilevel item response model. Four models were fit to the data.

The first model is the full multilevel latent regression item response model, specified as:

\[
\begin{align*}
\text{logit}(Y_{ipc} = 1 | \theta_{pc}) &= \theta_{pc} - \delta_i \\
\theta_{pc} &= X'_{pc} \beta + \xi_c + \epsilon_{pc} \\
\xi_c &\sim N(0, \psi_b) \\
\epsilon_{pc} &\sim N(0, \psi_w),
\end{align*}
\]

(2.47) (2.48) (2.49) (2.50)

where \(Y_{ipc}\) are the 24 dichotomous item responses and \(X'_{pc}\) contains the six covariates listed in Table 2.6. The results from this model served as the baseline. This is the correct standard in the absence of survey weights. Since the student level survey weights were non-informative in this example, this model was also treated as the baseline when the survey weights were used, though in that case it is an approximate standard.

The second model is the proposed multilevel model using WLEs as the outcome and accounting for measurement error,

\[
\begin{align*}
\hat{\theta}^{WL}_{pc} &= X'_{pc} \beta + \xi_c + \epsilon_{pc} + \zeta_{pc} \\
\xi_c &\sim N(0, \psi^{(2)}) \\
\epsilon_{pc} &\sim N(0, \psi^{(1)}) \\
\zeta_{pc} &\sim N(0, \sigma^2_{pc}) \\
\sigma^2_{pc} &= \text{var}(\hat{\theta}^{WL}_{pc}),
\end{align*}
\]

(2.51) (2.52) (2.53) (2.54) (2.55)

where \(\hat{\theta}^{WL}_{pc}\) is the WLE constructed from the 24 items and \(X'_{pc}\) contains the six covariates listed in Table 2.6.

The third and fourth models are plausible value analyses of the two separate sets of PVs, constructed with conditioning models similar to NAEP and PISA as discussed in Table 2.6:

\[
\begin{align*}
\theta^{PV}_{pc} &= X'_{pc} \beta + \xi_c + \epsilon_{pc} \\
\xi_c &\sim N(0, \psi^{(2)}) \\
\epsilon_{pc} &\sim N(0, \psi^{(1)})
\end{align*}
\]

(2.56) (2.57) (2.58)

where \(\theta^{PV}_{pc}\) is the \(m\)th PV constructed from the 24 item responses and \(X'_{pc}\) contains the six covariates listed in Table 2.6.

Each of the above models model was fit to the data twice, first without the survey weights and then with survey weights. When the survey weights were used, the standard errors were calculated in two ways: first, using the Sandwich estimator, and second, using the BRR
replicate weights and calculating the sampling variance using Equation 2.46. In addition, when the survey weights were used, they were scaled in three ways: raw (left unscaled), size, and effective.

### 2.4.3 Results

The results are presented in four sections corresponding to the following four research questions:

1. How does the proposed method compare to the full measurement model or the plausible value analysis?
2. How does the comparison between the methods change if the survey weights are used?
3. How does the comparison between the methods change if the standard errors are calculated using the replicate weights instead of the Sandwich estimator?
4. How does the comparison between the methods change if the student level weights are scaled differently?

#### Model comparison without weights

Table 2.7 gives the results from applying each of the four analysis models to the empirical data without the survey weights. Figure 2.13 graphically compares the estimated regression coefficients and their standard errors from each of the secondary analysis methods to the full measurement model.

Considering only the random part of the model, all three secondary analysis methods differ from the full measurement model. The differences match those found in the simulation studies. Under the proposed method, the student-level variance is severely underestimated and the school-level variance is slightly underestimated as well. The proposed method performs poorly in this example due to the small number of items whose difficulties are not so evenly spaced. For the PVS constructed as in NAEP (just using the covariates), the school-level variance is underestimated while the student-level variance is slightly overestimated. For the PVS constructed as in PISA (using the covariates and school dummies), the school-level variance is severely overestimated and the student-level variance is slightly underestimated. The PVS perform poorly in this example due to the very small number of students per school; in the full PISA sample, using the results from all of the booklets, the PVS would perform better. The different misspecifications of the conditioning models for the PVS either under- or over-account for the variation between the schools.

Considering only the fixed part of the model, all three secondary analysis methods perform similarly to each other and to the full multilevel measurement model. The estimated regression coefficients are similar across all four methods. These results correspond with the findings from the second simulation study that there was little bias in any of the regression
Table 2.7: Comparison of methods, without including survey weights.

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>WLEs</th>
<th>PVS (NAEP)</th>
<th>PVS (PISA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scmatedu</td>
<td>0.008</td>
<td>0.001</td>
<td>-0.004</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>(0.040)</td>
<td>(0.034)</td>
<td>(0.034)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>perfrlun</td>
<td>-0.004*</td>
<td>-0.003*</td>
<td>-0.003*</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>hedres</td>
<td>0.028</td>
<td>0.020</td>
<td>0.029</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.040)</td>
<td>(0.039)</td>
<td>(0.046)</td>
</tr>
<tr>
<td>female</td>
<td>-0.156*</td>
<td>-0.147*</td>
<td>-0.134*</td>
<td>-0.169*</td>
</tr>
<tr>
<td></td>
<td>(0.068)</td>
<td>(0.058)</td>
<td>(0.058)</td>
<td>(0.076)</td>
</tr>
<tr>
<td>race1</td>
<td>-0.415***</td>
<td>-0.378***</td>
<td>-0.425***</td>
<td>-0.485***</td>
</tr>
<tr>
<td></td>
<td>(0.102)</td>
<td>(0.087)</td>
<td>(0.090)</td>
<td>(0.086)</td>
</tr>
<tr>
<td>race3</td>
<td>-0.891***</td>
<td>-0.785***</td>
<td>-0.937***</td>
<td>-0.928***</td>
</tr>
<tr>
<td></td>
<td>(0.115)</td>
<td>(0.098)</td>
<td>(0.109)</td>
<td>(0.127)</td>
</tr>
<tr>
<td>race4</td>
<td>0.422*</td>
<td>0.322</td>
<td>0.429**</td>
<td>0.397*</td>
</tr>
<tr>
<td></td>
<td>(0.193)</td>
<td>(0.167)</td>
<td>(0.161)</td>
<td>(0.185)</td>
</tr>
<tr>
<td>race5</td>
<td>-0.211</td>
<td>-0.225</td>
<td>-0.157</td>
<td>-0.244</td>
</tr>
<tr>
<td></td>
<td>(0.176)</td>
<td>(0.150)</td>
<td>(0.160)</td>
<td>(0.194)</td>
</tr>
<tr>
<td>race6</td>
<td>-0.634*</td>
<td>-0.583**</td>
<td>-0.611*</td>
<td>-0.649**</td>
</tr>
<tr>
<td></td>
<td>(0.249)</td>
<td>(0.210)</td>
<td>(0.255)</td>
<td>(0.212)</td>
</tr>
<tr>
<td>escs</td>
<td>0.349***</td>
<td>0.303***</td>
<td>0.367***</td>
<td>0.342***</td>
</tr>
<tr>
<td></td>
<td>(0.048)</td>
<td>(0.041)</td>
<td>(0.045)</td>
<td>(0.041)</td>
</tr>
<tr>
<td>cons</td>
<td>0.030</td>
<td>0.011</td>
<td>0.092</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.071)</td>
<td>(0.071)</td>
<td>(0.095)</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(stu)</td>
<td>.858</td>
<td>.506</td>
<td>.892</td>
<td>.761</td>
</tr>
<tr>
<td></td>
<td>(.062)</td>
<td>(.044)</td>
<td>(.051)</td>
<td>(.041)</td>
</tr>
<tr>
<td>var(sch)</td>
<td>.060</td>
<td>.044</td>
<td>.021</td>
<td>.163</td>
</tr>
<tr>
<td></td>
<td>(.027)</td>
<td>(.020)</td>
<td>(.025)</td>
<td>(.041)</td>
</tr>
<tr>
<td>ICC</td>
<td>0.065</td>
<td>0.080</td>
<td>0.023</td>
<td>0.176</td>
</tr>
</tbody>
</table>
coefficients. The estimated standard errors show more variation but are still relatively similar across the methods. Using the proposed WLE method, the standard errors for all of the coefficients are smaller than the standard errors from the full measurement model. This is a consequence of the underestimated residual variance at both levels. Using the PVs, some of the standard errors are larger and some are smaller; the standard errors for the school-level covariates are underestimated for the PVs constructed as in NAEP and overestimated for the PVs constructed as in PISA while the standard errors for the student-level covariates show no discernable pattern. This is again a consequence of the residual variances being over- or underestimated.

The intraclass correlation was calculated for each model to consider one consequence of these differences in the estimated variance components. By underestimating both variance components, the proposed method best recovers the intraclass correlation from the full measurement model. The NAEP-style PVs underestimate the intraclass correlation while the PISA-style PVs overestimate it. Using any of the methods, caution should be exercised when making any interpretation that relies on the variance components estimates.

Model comparison with survey weights

Table 2.8 gives the results from applying each of the four analysis models to the empirical data using the final fixed survey weights scaled using the size method. Figure 2.14 graphically compares the estimated regression coefficients and their standard errors from each of the secondary analysis methods to the full measurement model. Though the inclusion of the survey weights in the analysis alters the estimated effects, and therefore any substantive

Figure 2.13: Comparison of the secondary analyses to the full measurement model without survey weights.
CHAPTER 2. MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

Figure 2.14: Comparison of the secondary analyses to the full measurement model with survey weights.

conclusions drawn with regard to the hypothetical empirical research questions, they do not alter any of the comparisons between the different analysis methods.

Considering the random part of the model, all three secondary analyses poorly estimate the variance components. The model with WLEs severely underestimates the student-level residual variance and slightly underestimates the school-level residual variance. The analysis with PVs constructed as in NAEP slightly overestimates the student-level residual variance and underestimates the school-level variance, approaching the boundary of the parameter at zero. The analysis with PVs constructed as in PISA greatly overestimates the school-level variance and underestimates the student-level variance. These differences are reflected in the estimated $\text{ICC}$.

Considering the fixed part of the model, the estimated regression coefficients are still similar across all four methods. The estimated standard errors show the same patterns of under- and overestimation as before. The greatest differences between the $\text{SEs}$ from the full measurement model and the $\text{SEs}$ from the secondary analyses are for the two PV analyses for the race dummies.

Based on these results, none of the secondary analysis methods are recommended above the other. Rather, in particular, the poor performance of all methods as compared to the simulation study emphasizes the importance of the cluster size and number of clusters and items in constructing estimators that are good enough for a secondary analysis.
Table 2.8: Comparison of methods, including fixed final survey weights.

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>WLEs</th>
<th>PVS (NAEP)</th>
<th>PVS (PISA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scmatedu</td>
<td>-0.005</td>
<td>-0.016</td>
<td>0.007</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.062)</td>
<td>(0.051)</td>
<td>(0.056)</td>
<td>(0.077)</td>
</tr>
<tr>
<td>perfrlun</td>
<td>-0.005*</td>
<td>-0.004*</td>
<td>-0.004</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>hedres</td>
<td>0.055</td>
<td>0.058</td>
<td>0.044</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>(0.094)</td>
<td>(0.079)</td>
<td>(0.088)</td>
<td>(0.081)</td>
</tr>
<tr>
<td>female</td>
<td>-0.037</td>
<td>-0.019</td>
<td>-0.050</td>
<td>-0.094</td>
</tr>
<tr>
<td></td>
<td>(0.109)</td>
<td>(0.098)</td>
<td>(0.099)</td>
<td>(0.118)</td>
</tr>
<tr>
<td>race1</td>
<td>-0.728***</td>
<td>-0.689***</td>
<td>-0.647***</td>
<td>-0.700***</td>
</tr>
<tr>
<td></td>
<td>(0.210)</td>
<td>(0.182)</td>
<td>(0.196)</td>
<td>(0.186)</td>
</tr>
<tr>
<td>race3</td>
<td>-1.240***</td>
<td>-1.037***</td>
<td>-1.143***</td>
<td>-1.178***</td>
</tr>
<tr>
<td></td>
<td>(0.202)</td>
<td>(0.153)</td>
<td>(0.166)</td>
<td>(0.258)</td>
</tr>
<tr>
<td>race4</td>
<td>0.414*</td>
<td>0.295</td>
<td>0.383</td>
<td>0.394</td>
</tr>
<tr>
<td></td>
<td>(0.197)</td>
<td>(0.167)</td>
<td>(0.210)</td>
<td>(0.235)</td>
</tr>
<tr>
<td>race5</td>
<td>-0.218</td>
<td>-0.241</td>
<td>-0.180</td>
<td>-0.069</td>
</tr>
<tr>
<td></td>
<td>(0.401)</td>
<td>(0.333)</td>
<td>(0.218)</td>
<td>(0.323)</td>
</tr>
<tr>
<td>race6</td>
<td>-0.860***</td>
<td>-0.835***</td>
<td>-0.791*</td>
<td>-0.864**</td>
</tr>
<tr>
<td></td>
<td>(0.117)</td>
<td>(0.098)</td>
<td>(0.339)</td>
<td>(0.294)</td>
</tr>
<tr>
<td>escs</td>
<td>0.298***</td>
<td>0.253***</td>
<td>0.309***</td>
<td>0.311***</td>
</tr>
<tr>
<td></td>
<td>(0.070)</td>
<td>(0.062)</td>
<td>(0.089)</td>
<td>(0.067)</td>
</tr>
<tr>
<td>cons</td>
<td>0.090</td>
<td>0.048</td>
<td>0.164</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.099)</td>
<td>(0.136)</td>
<td>(0.157)</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(stu)</td>
<td>0.766</td>
<td>0.388</td>
<td>0.811</td>
<td>0.711</td>
</tr>
<tr>
<td></td>
<td>(0.112)</td>
<td>(0.081)</td>
<td>(1.836)</td>
<td>(0.078)</td>
</tr>
<tr>
<td>var(sch)</td>
<td>0.036</td>
<td>0.029</td>
<td>0.000</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.021)</td>
<td>(637.819)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>ICC</td>
<td>0.045</td>
<td>0.070</td>
<td>0.000</td>
<td>0.166</td>
</tr>
</tbody>
</table>
CHAPTER 2. MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

Figure 2.15: Standard errors from BRRs versus the Sandwich estimator.

Replicate weights

Figure 2.15 displays a scatterplot of the standard errors for each regression coefficient comparing those from the Sandwich estimator to those calculated using the BRRs for each secondary analysis method. All of the points lie on or near the identify line, indicating that there are only minor differences observed in the estimated standard errors between the Sandwich estimator and the BRRs. This is a little surprising since the Sandwich estimator as implemented does not explicitly account for the stratification (though a Sandwich type estimator that does take this into account could be programmed).

As a result, there were no changes in the comparison of the different secondary analysis methods if the BRR standard errors are used instead of the Sandwich estimator. While no general conclusions can be drawn regarding the equivalence of these standard error methods, this analysis does suggest that further work should be done to determine when the use of BRR standard errors is necessary. It also shows that the BRR technique does not compensate for poor estimators of the latent variable. For example, if the PVs do not vary sufficiently between schools, as is the case with PVs constructed as in NAEP, resampling schools will not lead to sufficient variability for the parameter estimates and will still underestimate the standard errors.

Scaling of the weights

Table 2.9 gives the results from the full measurement model with the survey weights scaled in three different ways: (1) to sum to the actual cluster size (size), (2) to sum to the effective cluster size (effective), and (3) with no scaling (raw). Figure 2.16 graphically compares the estimated regression coefficients and their standard errors. The results from the three secondary analyses are not reported because the conclusions are similar.
Table 2.9: Comparison of different scaling methods for the student-level weights for the full measurement model.

<table>
<thead>
<tr>
<th></th>
<th>Size b/se</th>
<th>Effective b/se</th>
<th>Raw b/se</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scmatedu</td>
<td>-0.005</td>
<td>-0.005</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>(0.062)</td>
<td>(0.062)</td>
<td>(0.051)</td>
</tr>
<tr>
<td>perfrln</td>
<td>-0.005*</td>
<td>-0.005*</td>
<td>-0.006**</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>hedres</td>
<td>0.055*</td>
<td>0.055</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>(0.094)</td>
<td>(0.095)</td>
<td>(0.053)</td>
</tr>
<tr>
<td>female</td>
<td>-0.037</td>
<td>-0.037</td>
<td>-0.128</td>
</tr>
<tr>
<td></td>
<td>(0.109)</td>
<td>(0.109)</td>
<td>(0.071)</td>
</tr>
<tr>
<td>race1</td>
<td>-0.728***</td>
<td>-0.728***</td>
<td>-0.603***</td>
</tr>
<tr>
<td></td>
<td>(0.210)</td>
<td>(0.210)</td>
<td>(0.120)</td>
</tr>
<tr>
<td>race3</td>
<td>-1.240***</td>
<td>-1.241***</td>
<td>-0.974***</td>
</tr>
<tr>
<td></td>
<td>(0.202)</td>
<td>(0.202)</td>
<td>(0.129)</td>
</tr>
<tr>
<td>race4</td>
<td>0.414*</td>
<td>0.414*</td>
<td>0.329</td>
</tr>
<tr>
<td></td>
<td>(0.197)</td>
<td>(0.197)</td>
<td>(0.200)</td>
</tr>
<tr>
<td>race5</td>
<td>-0.218</td>
<td>-0.217</td>
<td>-0.449*</td>
</tr>
<tr>
<td></td>
<td>(0.401)</td>
<td>(0.402)</td>
<td>(0.201)</td>
</tr>
<tr>
<td>race6</td>
<td>-0.860***</td>
<td>-0.859***</td>
<td>-0.787***</td>
</tr>
<tr>
<td></td>
<td>(0.117)</td>
<td>(0.117)</td>
<td>(0.164)</td>
</tr>
<tr>
<td>7escs</td>
<td>0.298***</td>
<td>0.298***</td>
<td>0.301***</td>
</tr>
<tr>
<td></td>
<td>(0.070)</td>
<td>(0.070)</td>
<td>(0.049)</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(stu)</td>
<td>0.766</td>
<td>0.766</td>
<td>0.680</td>
</tr>
<tr>
<td></td>
<td>(0.112)</td>
<td>(0.112)</td>
<td>(0.061)</td>
</tr>
<tr>
<td>var(sch)</td>
<td>0.036</td>
<td>0.036</td>
<td>0.169</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td>(0.029)</td>
<td>(0.018)</td>
</tr>
</tbody>
</table>
CHAPTER 2. MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

Figure 2.16: Comparison of the different scaling methods for the student-level weights for the full measurement model.

The results for the size and effective scaling (first two columns) are identical, resulting in the dark circles and light diamonds being plotted on top of each other in Figure 2.16. This was expected since the student level weights were non-informative after being rescaled despite the small cluster sizes.

The results from the unscaled weights differ from the scaled weights. As expected from results in Pfeffermann et al. (1998) and Rabe-Hesketh & Skrondal (2006), school level variance was much larger (probably overestimated) and student level variance was smaller when using the raw weights. There are also differences in the estimated regression coefficients, which are largest for the race dummy variables, though there is no discernable pattern. In addition, the estimated ses are almost all smaller with the raw survey weights; again, the largest differences are observed for the race dummies.

These results confirm the importance of scaling the survey weights appropriately if one wants to use the analysis to answer a research question such as the hypothetical one used in this empirical study. It also suggests that in this particular data set, the analysis with survey weights does not have large bias despite the small cluster sizes.

2.5 Discussion

This paper proposes and assesses a method for (secondary) data analysis when the outcome variable in a multilevel model is latent and therefore measured with error. The proposed method accounts for the potentially non-constant error variance in the outcome variable by incorporating a classical measurement model with heteroskedastic variance at the lowest level.
CHAPTER 2. MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

of the multilevel model. The goal is a method that is convenient and readily understandable for applied research wherein a latent variable is being used as the response variable in a multilevel model.

PVs are the current standard for secondary analysis of a latent variable. This technique relies on having a correct conditioning model for the intended secondary analysis. PVs that are constructed with a single-level regression as the conditioning model do not perform well when the subsequent analysis is a multilevel model. Conditioning models for the PVs using only covariates to account for variance at the higher level, as is currently done in NAEP, result in biased estimates of the variance components for even 50 items. The inclusion of dummy variables to account for the clustering, as is currently done in PISA, overcompensates for this bias. This bias increases as the sample size and number of items decreases, as seen most clearly in the empirical study; with smaller clusters and fewer items, the variance components did not match the results of a full multilevel item response model. The incorrect estimation of the variance components leads to poor standard errors; for example, underestimation of the school-level variance leads to underestimating the standard errors for the school-level covariates. BRR resampling does not improve the standard errors. More than 25 items and 100 clusters with 50 people per cluster were needed for the bias to become negligible. In the full PISA sample from the United States, the school sizes and number of schools and items approaches these values.

The proposed method uses weighted likelihood estimates (WLEs; Warm, 1989) of the latent variable from fitting an item response model as the dependent variable for the multilevel model. It explicitly accounts for measurement error in the WLEs by fixing part of the level-1 residual variance equal to the estimated variance of the WLEs. By using WLEs, the proposed method does not require a correct conditioning model in order to recover the variance components well. This is particularly important since the actions of secondary data analysts are impossible to predict; for example, they may include variables from other data sources or may specify a variety of interactions or random coefficients. However, the method does rely on the precision of the estimation of the latent variable; if the latent variable is not precisely estimated, the person-level variance is underestimated.

The viability of the proposed method relies on the quality of the WLEs and their standard errors. The WLE was selected because the method requires a conditionally unbiased estimator and standard error of that estimator. Asymptotically, the WLE meets these criteria; in a finite sample, more than 25 items well matched to the range of the latent variable appears sufficient. 50 items covering the full range of the latent variable were required for unbiased estimates of the lowest-level variance component of the multilevel model; the regression coefficients and higher-level variance component could be recovered with fewer items. In large-scale assessments, such as in the PISA data examined here, students responded to no more than 24 items, resulting in poor recovery of the results of a full multilevel item response model. However, if targeted or adaptive tests were used instead, which is a real possibility as large scale assessments are becoming computerized, fewer items per person should be needed for the proposed method to work well.

Most large-scale assessments, for which secondary data analysis techniques are most
widely applied, require analysis that accounts for sampling design. The empirical study
gives only a first look at the use of survey and replicate weights in secondary data analysis
procedures such as the proposed method. In this particular case, neither affected the per-
formance of the secondary data analysis methods. However, further work on the effect of
incorporating information about the sampling design into the analysis is important because
none of the estimators considered in this paper, nor the actual PVs released by PISA (OECD,
2009a), consider the sampling design when constructing the estimators for secondary analy-
sis.

There are a number of additional avenues for further work in this area. We did not
consider the recovery of regression coefficients for interactions or random coefficients under
any of the estimators. In this paper, we examined only hierarchical linear models for the
latent variable. We could consider other uses of the estimators of the latent variable. For
example, sometimes the estimators are used to construct cutpoints, such as a proficiency
standard, that are subsequently used in a nonlinear multilevel analysis. We could explore
how the alternative secondary analysis methods perform for these models. In this paper, we
examined only a unidimensional latent variable. It is common in large-scale assessment to
address a multidimensional set of latent variables. PVs could be created that account for the
multidimensionality (as is currently done in PISA and NAEP); the WLEs used in the proposed
method would not borrow information across dimensions. We could explore the performance
of the PVs in univariate analyses if they were constructed accounting for multidimensionality
and the performance of both estimators in multivariate analyses in either case.

The current paper explicitly set out to examine a method for secondary analysis under
the assumption that an alternative to PVs was needed for models with a random intercept.
The implementation of a viable method for constructing PVs that accounts for this aspect
of hierarchical structure could obviate this concern. However, it is difficult to predict the
full complexity of the models that secondary users will fit to any data. Possibilities include
interactions between variables, random coefficients, or even merging with data from another
source to explore additional variables. It is an impossible task to construct PVs that are
appropriate for all possible subsequent analyses. The proposed WLE method is therefore
promising since it does not require any assumptions about the models used in secondary
data analysis.
2.A Appendix

2.A.1 Model syntax for Simulation 1

The estimated models for the first simulation are written using Stata syntax as:

```
/*
Data contains the following variables:
   clusid -- cluster id number
   wle -- weighted likelihood estimate of latent variable
   wleerr -- standard error of wle
   theta -- simulated true value of the latent variable
*/

**proposed model: accounting for error in wles using SEs
*transform wle errors to estimation scale (variance)
gen wlevar = wleerr^2
gen cons = 1

*define equation for level 1 errors
eq het: cons wlevar

*define linear constraint
constraint def 1 [s2]wlevar = 1

*matrix of starting values
matrix a = (0, .2, 1, .5)

*fit model: use s() and constr() options and natvar with starting values
gllamm wle, i(clusid) s(het) constr(1) adapt natvar from(a) copy long

**naive model: ignoring error in wles
*estimate using xtmixed
xtmixed wle || clusid:

**simulated abilities model: using true thetas
*estimate using xtmixed
xtmixed theta || clusid:
```
2.A.2 Model syntax for Simulation 2

The estimated models for the second simulation including both covariates are written using Stata syntax as:

/*
Data contains the following variables:
  clusid -- cluster id number
  wthn -- covariate that varies only within clusters
  btwn -- covariate that varies only between clusters
  wle -- weighted likelihood estimate of latent variable
  wleerr -- standard error of wle
  pv1-pv5 -- 5 plausible values for the latent variable
  theta -- simulated true value of the latent variable
*/

**proposed model: accounting for error in wles using SEs
*transform wle errors to estimation scale (variance)
gen wlevar = wleerr^2
gen cons = 1

*define equation for level 1 errors
eq het: cons wlevar

*define linear constraint
constraint def 1 [s2]wlevar = 1

*matrix of starting values
matrix a = (0, 0, 0, .2, 1, .5)

*fit model: use s() and constr() options and natvar with starting values
gllamm wle wthn btwn, i(clusid) s(het) constr(1) adapt natvar ///
    from(a) copy long

**naive model: ignoring error in wles
*estimate using xtmixed
xtmixed wle wthn btwn || clusid:

**PV model (code repeated for different sets of pvs)
*mi set the data
  gen pv = .
  mi import wide, imputed(pv = pv1 pv2 pv3 pv4 pv5) drop clear

*fit model
  mi estimate, dots post: xtmixed pv wthn btwn || clusid:,

  **simulated abilities model: using true thetas
  *estimate using xtmixed
  xtmixed theta btwn || clusid:
Chapter 3

Longitudinal explanatory item response models to evaluate educational interventions

3.1 Introduction

Educational interventions are designed based on a specific theory of learning and development (e.g. the intervention’s logic model or theory of action). Ideally, we would want to consider these design aspects when conducting an evaluation of the intervention. This would simultaneously give more nuanced information regarding how the program was effective as well as information regarding the soundness of the developmental theory. This paper proposes longitudinal, explanatory item response models for this purpose.

For example, in the Learning Mathematics through Representations (LMR) project (Saxe et al., 2013), two important design principles of the new lesson sequence are (1) the use of the number line as the central mathematical representation and (2) the joint construction of a set of principles and definitions. These are intended to build coherence for students to extend their understanding of number from whole numbers to integers to fractions. The assessments then include items that use different representations and tap into different principles across two content domains. We might expect that initial learning gains on the assessments would be driven by student familiarity with the number line but that longer term learning gains would not as students generalized their knowledge to other problems. We could test this hypothesis by comparing the relative difficulty of items with and without number lines over time. Thus, by modeling item features such as the presence of a number line, we would be able to make more nuanced explanations about the efficacy of the LMR curriculum.
3.1.1 Previous models for explanatory and longitudinal purposes

The linear logistic test model (LLTM; Fischer, 1973, 1983) has been used previously to model cognitive features and estimate the effects of item features on the difficulty of items. A basic item response model, the Rasch model (Rasch, 1960/1980), can be specified as:

\[
\text{logit}(\Pr(Y_{ip} = 1 | \theta_p)) = \theta_p - \delta_i, \tag{3.1}
\]

where \(\delta_i\) is the difficulty of item \(i\), \(\theta_p\) is the proficiency of person \(p\) (often assumed \(\sim N(0, \psi)\)), and \(Y_{ip}\) is a response to an item that is scored 1 if correct and 0 if incorrect. The LLTM modifies this model by using Equation 3.1 but modeling the item difficulties as:

\[
\delta_i = \sum_{k=0}^{K} \beta_k X_{ki}, \tag{3.2}
\]

where \(\beta_k\) is a parameter for the effect of an item covariate \(X_{ki}\). Usually, the \(X_{ki}\) are indicators (i.e. coded as 0 or 1) of item features which correspond to the presence or absence of factors that are hypothesized to affect the item difficulty (though other types of covariates are possible, depending on the available theory). For example, in the LMR data, one such item covariate would be whether the item used a number line or not. In general, each item has multiple item features and thus multiple item covariates are included to model item difficulty. The use of more item covariates also ensures a better explanation of the item difficulty.

The framework of the LLTM was extended to model longitudinal growth and treatment effects by incorporating fixed main effects and interactions for time and treatment group into the item difficulty side of the item response model (Fischer, 1989; Gluck & Spiel, 1997; Kubringer, 2008). When the item difficulties were allowed to change over time (using interactions), the interpretation was in terms of measurement invariance and possible multidimensionality.

Extensions to the LLTM to make it more flexible and allow it to better fit data include allowing a random residual for the items (random intercept LLTM; Janssen, Schepers & Peres, 2004; Janssen, 2010) and random coefficients of item covariates (random weights LLTM; Rijmen & De Boeck, 2002). Prowker & Camilli (2007) extended the random intercept LLTM by including a higher-level random item intercept (i.e. item- and state-specific) to test for item difficulty variation across states. Similarly, Park & Bolt (2008) proposed an extension of the random-weights LLTM (though they did not frame it as such) to model differential effects of item features across countries.

Under the LLTM framework, changes in the persons’ proficiencies over time were modeled as changes in item difficulty. Models that directly parameterize change over time in the person part of the item response model were proposed by Andersen (1985) and Embretson (1991) using different parameterizations of the random person effects. Latent growth item response models (Zheng, 2009; McGuire, 2010; Wilson et al., 2012) built upon these models by taking advantage of the flexible measurement model machinery afforded by the multidimensional random coefficients multinomial logit (MRCML) model framework (Adams et al., 1997).
Davier et al. (2011) presents another overarching model, explicitly comparing the approaches of Andersen and Embretson for parameterizing the changes in person proficiency, and also adding extensions to accommodate manifest and latent person groups. Rijmen et al. (2005) also considered an extended item response model for group-specific person change over time by incorporating a latent transition model for person states into the item response model\footnote{This is similar to how person group differences are parameterized by the Saltus model (Wilson, 1989), which is a sub-model.}

Combining elements from the longitudinal models mentioned above, Pastor & Beretvas (2006) presented a longitudinal Rasch model as a hierarchical generalized linear model that allowed both person proficiency and item difficulty to change over time. Their model did not explicitly include item covariates. Note that in order for the model to be identified, at least one item difficulty must be fixed to be constant across time.

### 3.1.2 Purposes of this paper

This paper proposes and illustrates a longitudinal explanatory item response model that simultaneously models change between groups over time for persons and items. It involves extensions to an LLTM that allow the effects of item covariates to change over time (i.e. includes interactions of item covariates with time) and also over manifest groups (e.g. treatment conditions, which might be at a higher level such as classroom) in a three-way interaction. The proposed model is an extension of the model discussed by Pastor & Beretvas (2006) but with item covariates rather than item indicators\footnote{The $\delta_i$ in the Rasch model are the effects of item indicators, or dummy variables, whereas the $\beta_k$ in the LLTM are the effects of item covariates, or more accurately, the indicators for item covariates.} parameterizing item difficulty and with different effects of item covariates between treatment groups.

The purpose of the proposed models is to assess the relative change in the difficulty of item features over time and potential differential change over time due to group membership after controlling for overall person growth. Statistically, the proposed models are straightforward extensions; conceptually, the proposed models allow for a more fine-grained, explanatory analysis of factors that may contribute to change over time. In particular, when the change is thought to be due to a developmentally-motivated intervention, the proposed models allow us to test how item features related to the theory of action of the intervention may contribute to change over time.

The main contribution of this paper is to the theory and interpretation of time-varying item feature effects. I discuss and illustrate how explanatory, longitudinal item response models could be used to move beyond an evaluation of overall student gains to make more detailed inferences. Though similar substantive conclusions can be reached using other methods, such as analysis that combines summative assessment information with information gleaned from videos of classroom instruction, the utility of the proposed method lies in its ability to make more detailed inferences regarding the developmental aspects of the intervention from well-designed assessment data.
In Section 3.2, the proposed models are explicated with attention to issues regarding the link between identification, measurement invariance, and interpretation. In Section 3.3, the proposed models are applied to data from the LMR project, with a focus on challenges to interpretation given particular modeling choices. Three issues are examined: (1) how applying the model to a set of items that is well-described by the intervention and item design features can be used to explain the efficacy of an intervention, (2) how applying the model to a set of items that is less well-described by the design features affects its use, and (3) how applying the model with different identifying constraints impacts interpretation. The paper concludes with a discussion.

### 3.2 Longitudinal explanatory models for evaluation

#### 3.2.1 Model specification

The proposed models are defined sequentially using common notation. In the following, \( i \) indexes items, \( k \) indexes item covariates, \( t \) indexes occasions, and \( p \) indexes persons. \( Y_{iptp} \) is the response to item \( i \) at time \( t \) by person \( p \); it is scored 1 if correct and 0 if incorrect. \( X_{ki} \) is an item feature. \( z_{pt} \) is the time at occasion \( t \) for person \( p \) and \( g_{p} \) is the time-invariant group membership of person \( p \), taking the value 1 for the group of interest (e.g. the treatment group) and 0 for the reference group (e.g. the comparison group).

The Rasch model (Equation 3.1) can be extended to incorporate longitudinal effects as well as variation in item difficulty between people, as follows:

\[
\text{logit}(\Pr(Y_{iptp} = 1 \mid \theta_{tp})) = \theta_{tp} - \delta_{iptp}.
\] (3.3)

Note that now the person parameter \( \theta_{tp} \) is indexed by both \( t \) and \( p \) and the item parameter \( \delta_{iptp} \) is indexed by \( i \), \( t \), and \( p \). This is the base model (i.e. the “level 1 model”) that will be elaborated by specifying models for both \( \theta_{tp} \) and \( \delta_{iptp} \) to consider changes in person ability and changes in item difficulty over groups of persons over time. We will refer to \( \theta_{tp} \) and its model as the person side and to \( \delta_{iptp} \) and its model as the item side.

A typical model for \( \theta_{tp} \) in an evaluation context is a growth curve model, such as a model with fixed effects of group and time and a random intercept for people:

\[
\theta_{pt} = \pi_{p0} + \pi_{1}g_{p} + \pi_{2}z_{pt} + \pi_{3}z_{pt}g_{p} \quad (3.4)
\]

\( \pi_{p0} \) is the random intercept, giving person-specific residuals from the average growth curve. The mean is zero because the item difficulties are not constrained to sum to 0. \( \pi_{1} \) is the average difference between groups at baseline. \( \pi_{2} \) is the average growth rate for the reference group. \( \pi_{3} \) is the difference in growth rate between the focal and reference groups. There are many possible extensions to this model, for example considering a random coefficient for time, non-linear effects of time, or other person covariates. The models applied in this paper use the growth curve model presented here.
Models for variation in item difficulty

Variations in how $\delta_{itp}$ is modeled reflect the different components of the final, proposed model.

The simplest model for $\delta_{itp}$ assumes that it does not change over time or differ between groups:

$$\delta_{itp} = \gamma_{i00}. \quad (3.5)$$

$\gamma_{i00}$ is the item difficulty. This model (Equations 3.3, 3.4, and 3.5) is a latent regression Rasch model (Mislevy, 1987; Adams et al., 1997; Kamata, 2001; Zwinderman, 1991; Rijmen et al., 2003; Van den Noortgate & Paek, 2004).

If $\delta_{itp}$ is allowed to vary between people, we can model how the item difficulty differs between the groups of people being evaluated:

$$\delta_{itp} = \gamma_{i00} + \gamma_{i01}g_p. \quad (3.6)$$

$\gamma_{i01}$ is the difference in item difficulty between groups. This model (Equations 3.3, 3.4, and 3.6) corresponds to IRT-based models for differential item functioning (DIF; Muthén, 1988; Thissen, Steinberg & Wainer, 1993; Meulders & Xie, 2004) but with a more complex latent regression.

If $\delta_{itp}$ is allowed to vary between occasions, we can model how the item difficulty changes over time:

$$\delta_{itp} = \gamma_{i00} + \gamma_{i10}z_{pt}. \quad (3.7)$$

$\gamma_{i10}$ is the change in item difficulty per unit of time. This model (Equations 3.3, 3.4, and 3.7) corresponds to the hierarchical generalized linear model for repeated measures of Pastor & Beretvas (2006).

If $\delta_{itp}$ is allowed to vary both between people and over time, we can model both simultaneously:

$$\delta_{itp} = \gamma_{i00} + \gamma_{i01}g_p + \gamma_{i10}z_{pt} + \gamma_{i11}z_{pt}g_p. \quad (3.8)$$

$\gamma_{i11}$ is the difference between groups in the change in item difficulty per unit of time. This model (Equations 3.3, 3.4, and 3.8) is a latent regression Rasch model with time-varying, differential item effects. This model provides a first step in exploring how person growth coexists with item changes. To ease interpretation, the coefficients $\gamma_{irs}$ have subscripts $r$ for covariates related to time and $s$ for covariates related to person groups.

Models to explain variation in item difficulty

The above models describe differences in item difficulty between groups of people over time. However, any variation that is found can only be explained as part of a second-stage analysis.
The following models directly incorporate one way to explain item changes. De Boeck & Wilson (2004) refer to models that describe differences in item difficulty as “descriptive on the item side” and to models that incorporate covariates to explain differences in item difficulty as “explanatory on the item side”. Since the models in this paper are all explanatory on the person side (i.e. they all include covariates for time and group membership to model $\theta_{tp}$), we use the terms descriptive and explanatory to refer to the differences on the item side.

Consider a set of item features $X_{ki}$ that are expected to explain item difficulty:

$$\delta_{itp} = \sum_{k=1}^{K} \beta_{ktp}X_{ki}. \tag{3.9}$$

The effects $\beta_{ktp}$ of item features can change over time $t$ and between persons $p$. As above, variations in how $\beta_{ktp}$ is modeled reflect different components of the final, proposed model.

Note that, as written, Equation 3.9 assumes that the specified item features perfectly explain the item difficulty. This assumption can be relaxed by including a random residual for the items (Mislevy, 1988; Janssen et al., 2004; Janssen, 2010):

$$\delta_{itp} = \sum_{k=1}^{K} \beta_{ktp}X_{ki} + \epsilon_i, \tag{3.10}$$

$$\epsilon_i \sim N(0, \tau). \tag{3.11}$$

The inclusion of a random effect for the items makes this a cross-classified model (Goldstein, 1987; Raudenbush, 1993); in psychometrics, models with crossed random effects for persons and items are called random item models (Van den Noortgate, De Boeck & Meulders, 2003; De Boeck, 2008).

In general, estimates of $\beta_{ktp}$ from a model without a random residual for items will be attenuated and their standard errors will be underestimated (Janssen, 2010). The standard errors are underestimated because the model does not account for within-cell variance of the item features. The severity of the attenuation is inversely proportional to $\tau$. To understand why, consider the model defined by Equations 3.3 and 3.10 written in latent response format:

$$Y_{itp} = I[Y_{itp}^* > 0], \tag{3.12}$$

$$Y_{itp}^* = \theta_{tp} - \sum_{k=1}^{K} \beta_{ktp}X_{ki} + \epsilon_i + \epsilon_{itp}, \tag{3.13}$$

$$\epsilon_i \sim N(0, \tau), \tag{3.14}$$

where $\epsilon_{itp}$ follows the standard logistic distribution. If $\epsilon_i$ is left out of the model, any item level residual variance $\tau$ will be absorbed into the $\epsilon_{itp}$. However, since the variance of $\epsilon_{itp}$ is fixed at $\frac{\pi^2}{3}$ for model identification (in logistic regression, only the ratio between the regression coefficients and the variance of the logistic distribution is identified), the “absorption” actually results in attenuation of the $\beta_{ktp}$. This phenomenon is referred to as
shrinkage in the measurement scale and will be important to keep in mind when applying these models.

The simplest model for $\beta_{ktp}$ assumes that it does not change over time or differ between groups of people:

$$\beta_{ktp} = \gamma_{k00}. \quad (3.15)$$

$\gamma_{k00}$ is now the effect of item feature $k$. This model (Equations 3.3, 3.4, 3.9, and 3.15) is a linear logistic test model (LLTM; Fischer, 1973) with a latent regression.

If $\beta_{ktp}$ is allowed to differ between groups of people, the model for $\beta_{ktp}$ becomes:

$$\beta_{ktp} = \gamma_{k00} + \gamma_{k01}g_p. \quad (3.16)$$

$\gamma_{k01}$ is the difference in the effect of item feature $k$ between groups. This model (Equations 3.3, 3.4, 3.9, and 3.16) is an IRT-based model for differential facet functioning (DFF; Engelhard, 1992; Meulders & Xie, 2004) with a latent regression.

If the effects of the item features are allowed to change linearly over time, the model for $\beta_{ktp}$ becomes:

$$\beta_{ktp} = \gamma_{k00} + \gamma_{k10}z_{pt}. \quad (3.17)$$

$\gamma_{k10}$ is the change in the effect of item feature $k$ per unit of time. This model (Equations 3.3, 3.4, 3.9, and 3.17) is a latent regression LLTM with time-varying item feature effects.

If $\beta_{ktp}$ is allowed to change both over time and differ between groups of people, we can model both simultaneously:

$$\beta_{ktp} = \gamma_{k00} + \gamma_{k01}g_p + \gamma_{k10}z_{pt} + \gamma_{k11}z_{pt}g_p. \quad (3.18)$$

$\gamma_{k11}$ is the difference between groups in the linear change over time of the item feature effect. This model (Equations 3.3, 3.4, 3.9, and 3.18) is a latent regression LLTM with time-varying, differential item feature effects. The model includes three-way interactions between occasion, group membership, and item feature. To ease interpretation, the coefficients $\gamma_{krs}$ have subscripts $r$ for covariates related to time and $s$ for covariates related to person groups.

This is the proposed model. Hypotheses regarding the $\gamma$s could be based on the design of the intervention being evaluated. In particular, it will probably often be assumed that item feature effects change only for the treatment group but not the comparison group for features tied to the intervention. These hypotheses, that certain features should show more relative change over time for certain groups if the intervention is working as anticipated, could be tested to explain the efficacy of the intervention.

### 3.2.2 Identification and interpretation

The models as written above are not identified due to the scale indeterminacy between the person-side and item-side once both $\theta$ and $\delta$ are allowed to vary over time and/or between persons. In order to identify the models, we need to constrain something to be constant both between groups and across time. There are two main possibilities for this constraint:
1. Constrain a single item difficulty or item feature effect.

2. Constrain the average of the item difficulties or item feature effects.

Usually, the same constraint is applied both across time and between groups. In practice, it is common to constrain more than one item difficulty to be constant; since this alters the model being estimated, in the following, we will discuss the constraint in reference to a single item.

Statistically, these two possibilities merely correspond to different ways of parameterizing the models. If dummy coding is used for the items and item features, the reference item difficulty or item feature effect is implicitly being constrained to be constant across time and groups. If effect (or contrast) coding is used for the items and item features, the average effects are implicitly being constrained to be constant across time and groups. The parameterization does not alter the statistical model being estimated. From a statistical point of view, the selection of the constraint should be innocuous.\(^3\)

However, from a theoretical or substantive point of view, the selection of a constraint and parameterization is not trivial. The first issue with the choice is related to the interpretation of the person growth latent regression. The choice of constraint alters the referent on the item side for the growth effects and therefore changes the meaning of any observed person growth. The second issue with the choice of parameterization is related to measurement invariance. The meaning of the scale will potentially change as the item parameters change, and the meaning of this change for subsequent interpretation will differ depending on the constraint.

Measurement invariance is defined to hold when the distribution of an observed random variable (e.g. an item response or test score) conditional on a latent variable of interest is the same for different groups or different occasions (Mellenbergh, 1989; Meredith & Millsap, 1992; Meredith, 1993). In the context of item response models, this translates into having the same item response function for different groups and different occasions. According to this definition, a lack of measurement invariance subsumes both differential item functioning (lack of measurement invariance between groups) and item parameter drift (lack of measurement invariance over time). Any item response function whose parameters (potentially) differ between groups or across occasions (as in the proposed model) does not meet this definition of measurement invariance.

This is not the only definition of measurement invariance. Meredith (1993) also defines a weaker condition, called weak measurement invariance, wherein only the first two moments (expectation and variance) of the conditional distribution of the observed variable are the same for different groups or different occasions. Similarly, the notion of a series of components of measurement invariance is well-established within the framework of confirmatory

---

\(^3\)There is a potential statistical issue from the perspective of identifying measurement invariance, regarding the difference between absolute and relative effects (e.g. is a single item showing large DIF versus the remaining items each showing slight DIF), that is not relevant for the purpose of the models being considered here.
factor analysis (see Vandenberg & Lance (2000) for a review of the literature in this area). These aspects of measurement invariance include “configural invariance” (in IRT, this corresponds to no variations in dimensionality), “metric invariance” (in IRT, no variations in item discrimination parameters), and “scalar invariance” (in IRT, no variations in item difficulties). An additional aspect arises when partial invariance is tested and controlled for if full invariance for any aspect is rejected. In their review, Vandenberg & Lance (2000) conclude that the use of measures with only partial invariance may be justified depending on the context and the theory behind the relaxation of invariance constraints.

The ideas of partial and weak measurement invariance open the possibility of considering whether there is an alternative definition of “measurement invariance” that is applicable to the proposed model. There are two questions that guide the consideration of a weaker but still useful finding of “measurement invariance”:

1. What is defined to be the same for different groups or occasions?
2. Is this definition sufficient to justify any desired comparisons?

If the constraint is placed on specific item difficulties or item feature effects, then (strong) measurement invariance is achieved but only for that subset of items. The meaning of the measurement scale is fixed, but any interpretation of change over time or differences between groups is restricted to a (limited) construct defined by a subset of the items. This does not seem to be sufficient to justify any desired conclusions regarding the larger intended construct. In contrast, if the constraint is placed on the average of the item difficulties or feature effects, then (strong) measurement invariance is achieved only for a hypothetical item representative of the average difficulty of the items. The meaning of the measurement scale is fixed by reference to the collection of items with that average difficulty, though the measurement scale in a more standard sense (that defined by the set of actual items) can change over time.

Constraining the average difficulty or feature effect is similar to how the identification constraint is applied in random item models for differential item functioning (DIF). The focus is on variation in item difficulty rather than identifying specific items that function differently. DIF is modeled using either a random coefficient for the interaction between item and person group indicators (treating groups as fixed but items as random; Van den Noortgate & De Boeck, 2005; Park & Bolt, 2008), a random interaction term (treating both groups and items as random; Van den Noortgate & De Boeck, 2005; Prowker & Camilli, 2007), or separate random item distributions within each group (De Boeck, 2008; de Jong & Steenkamp, 2010). Under all of these approaches, the identification constraint applied is to set the mean item difficulty to 0 within each group. Applying this constraint establishes a link between the item parameters without having to define a set of anchor items (De Boeck, 2008; de Jong & Steenkamp, 2010). De Boeck (2008) describes this as “posterior anchoring”,

---

4Park & Bolt (2008) applied the constraint to individual item features when estimating the model, but they then normalized the results (i.e. calculated what the results would have been had the average been constrained to zero) before interpreting them.
with the anchor set of items being the (latent) set of all non-DIF items. This constraint and the rationale is consistent with the constraint advocated for the proposed model, though the constraint in random item DIF is on the mean of the theoretical population of items rather than on the mean of the sample of items.

Determining the construct using the average of the item difficulties or item feature effects is also similar in spirit to both domain-level (Schulz, Lee & Mullen, 2005) and market-basket reporting (Mislevy, 1998; Mislevy & Zwick, 2012) for achievement over time. In domain-level reporting, growth results are reported in terms of an expected score (or set of scores) to a set of items chosen as representative of a domain for theoretical reasons (rather than based on their item statistics). The goal is to create aggregated score characteristic curves that do not cross even if the individual item characteristic curves do cross. Similarly, market-basket reporting gives results in terms of the expected score on a specified collection of tasks (not necessarily responded to by the individuals) chosen to create a single common scale for reporting across occasions and groups. Any results depend on the composition of the basket. In both of these approaches, information is combined or collapsed across items to create the measurement scale. This is similar to how constraining on the average difficulty or feature effect collapses information from the set of items into a single “average item” that is used to define the measurement scale.

The issue of how the choice of constraint alters the use and interpretation of the proposed model will be further illustrated in the empirical application.

### 3.2.3 Relation to other models

On the item side, the proposed model is closely related to previous IRT-based models for measurement non-invariance both (1) between groups (i.e. differential item functioning (DIF) and differential facet functioning (dff)) and (2) over time (i.e. item parameter invariance or drift (ipd)). The model specification where the item parameters differ only between groups or over time exactly corresponds to previous models for DIF/dff (Meulders & Xie, 2004; Paek & Wilson, 2011) and IPD (Pastor & Beretvas, 2006). However, the proposed models differ from other, more traditional models for DIF and IPD (e.g. Holland & Wainer, 1993; Donoghue & Isham, 1998, respectively), in that the focus is not on detecting items that violate measurement invariance but on explaining and interpreting any such findings. In particular, this shift in focus means that relative rather than absolute differences are of interest. The proposed model also only considers a particular form of measurement non-invariance.

On the person side, the proposed model directly follows from previous work for longitudinal item response models (Andersen, 1985; Embretson, 1991; von Davier et al., 2011; Wilson et al., 2012) and is closest to the model of Pastor & Beretvas (2006). In particular, the person side of the model as written has a more complicated fixed growth structure than previous models, modeling group membership, but a more restrictive random growth structure, containing only one random person effect. However, the person side could be extended to include more complicated growth models, such as reflected by the literature on latent
growth item response models. The person side could also be extended to include multidimensional person effects, such as Cho, Athay & Preacher (2013) did in their explanatory longitudinal growth model. The model of Cho et al. (2013) is most similar to the model of Equation 3.15 with the addition of (1) multivariate random effects for person to capture multidimensionality and (2) a random effect for items (Janssen, 2010) to capture variance unexplained by the item features.

The proposed model is a particular extension that combines elements from item response models for longitudinal and non-invariant data. The proposed model involves the simultaneous consideration of longitudinal person growth and two kinds of measurement non-invariance, between groups and over time. This is done with a particular purpose in mind, namely using differential relative changes in item features to explain differential person growth.

3.3 Application: Learning Mathematics through Representations (LMR)

The purpose of this study is to demonstrate the use of the proposed model with time-varying, differential item feature effects to evaluate an educational intervention. Data from an efficacy study of a new mathematics curriculum for elementary school is used to (1) determine if the proposed model can explain the efficacy of an educational intervention with a set of items that is well-described by the design features, (2) investigate how the proposed model functions when applied to a set of items that is less well-described by the design features, and (3) examine the link between identifying constraint and model interpretation.

3.3.1 Description of the LMR data

The Learning Mathematics through Representations (LMR) project has developed a supplemental curriculum for integers and fractions in the late-elementary grades. The curriculum consists of two two-week sequences of instruction, one for integers (9 lessons) and one for fractions (10 lessons). Two central features of the curriculum are (1) the use of the number line as a single central representation and (2) the use of a series of principles and definitions that reflect the core ideas of the curriculum and are explicitly discussed and agreed upon by the students.

An evaluation of the curriculum was conducted in 2010-2011 in 11 LMR and 10 Comparison classrooms in the San Francisco Bay Area (Saxe et al., 2013). The data analyzed for this example comes from that study. The sample considered here consists of item responses from 571 4th and 5th grade students to three assessments: a pretest in mid-September, a posttest in early December, and a final end-of-year test in mid-May. The same test form was used for the posttest and final test.

The assessment data being analyzed contained items that were the same on all three assessments (17 common items) and items that were unique to each test form (11 non-common
items on the pretest and 13 non-common items on the post/final test). Example common items are shown in Figure 3.1 and example non-common items are shown in Figure 3.2. The assessments were designed such that about half of the items directly correspond to tasks from the LMR curriculum and half are drawn from other sources, primarily the regular math textbook (which is the same in all classrooms). Dichotomous scoring was used for all items.

The items were designed around three primary item features (reflected in Figures 3.1 and 3.2):

- Item content: Integers or fractions
- Representation: Number line or not
- Response format: Constructed response (CR) or multiple choice (MC)

These three item features will be used as explanatory features in the models. The first two item features directly correspond to design features of the LMR curriculum, namely the number line as the central representation and the coherent extension of number line principles from integers to fractions. These are the item features for which we would expect to see differential time-varying effects if the LMR curriculum was effective due to its design principles.

The full set of items for the LMR assessments was systematically designed around these item features. However, other features also varied less systematically among the items, such as whether the item contained unit versus multiunit intervals (i.e. counting by 1s versus 2s, 5s, or 9s) or whether the item contained familiar versus unfamiliar subunit intervals (i.e. halves, thirds, and fourths versus sevenths, tenths, and fiftieths). When the common items were selected from the full item set, items were purposefully selected that best reflected all combinations of the systematic design features (i.e. integers and fractions, number line and non number line, multiple choice and constructed response). The intention was for the common items to be representative of the full set of items.

However, as seen in Figures 3.1 and 3.2, the common items are much better described by the design features than the non-common items. The figures show two items for each of the four cells defined by the content (integers or fractions) and representation (number line or not) item features. The items that are shown are those that had the lowest and highest estimated item difficulty in the efficacy study (Saxe et al., 2013); they are the easiest and hardest items in each cell. For the common items, the largest range in item difficulty within a cell is 3.3 logits for fractions non number line items and two cells (integers non number line items and fractions number line items) have ranges in item difficulty of less than 1 logit. In contrast, for the non-common items, the smallest range in difficulty within a cell is 1.3 logits for the integers non number line items and two cells (integers number line items and fractions non number line items) have ranges in item difficulty of almost 5 logits. The non-common items contain more variation in other item features that were not explicitly part of the item design, such as unusual multiunits and subunits. As a result, the common items are analyzed to illustrate the use and interpretation of the proposed models while the full
### (a) Integers items

<table>
<thead>
<tr>
<th>Number Line Task</th>
<th>Estimated Difficulty</th>
<th>Non Number Line Task</th>
<th>Estimated Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Easiest</strong></td>
<td>Mark with an arrow (↑) where 0 belongs on the number line.</td>
<td>1.1</td>
<td>Fill in the box with the correct sign (&gt;, &lt;, =).</td>
</tr>
<tr>
<td></td>
<td>-1.3</td>
<td>-1.0</td>
<td></td>
</tr>
<tr>
<td><strong>Hardest</strong></td>
<td>Write the number that belongs in the box.</td>
<td>2.3</td>
<td>What are the three missing numbers in this pattern? Write each missing number in the blank.</td>
</tr>
<tr>
<td></td>
<td>5, ___, ___, 18, ___</td>
<td>-0.3</td>
<td></td>
</tr>
</tbody>
</table>

### (b) Fractions items

<table>
<thead>
<tr>
<th>Number Line Task</th>
<th>Estimated Difficulty</th>
<th>Non Number Line Task</th>
<th>Estimated Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Easiest</strong></td>
<td>Write the number that belongs in the box.</td>
<td>1.9</td>
<td>Which fraction represents the largest part of a whole? Circle the answer.</td>
</tr>
<tr>
<td></td>
<td>0, 1</td>
<td>-1.7</td>
<td></td>
</tr>
<tr>
<td><strong>Hardest</strong></td>
<td>Write the number that belongs in the box.</td>
<td>2.3</td>
<td>What is the length of the toothpick in the figure below? Be as exact as you can.</td>
</tr>
<tr>
<td></td>
<td>0, 1, 2</td>
<td>1.6</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.1: Example common items from both representation and content domains; the easiest and hardest item is shown, with the estimated item difficulty (in logits) from Saxe et al. (2013).
Figure 3.2: Example non-common items from both representation and content domains; the easiest and hardest item is shown, with the estimated item difficulty (in logits) from Saxe et al. (2013).
set of items (common and non-common) are analyzed to explore the impact on the proposed models of excluding potentially important item features.

### 3.3.2 Research questions and analysis

The analysis using the LMR data is organized around three research issues:

1. Application of the proposed explanatory model to a set of items that is well-described by the item features to determine if the proposed model can explain the efficacy of an educational intervention.

2. Application of the proposed model to a set of items that is less well-described by the item features to investigate how this affects the use of the proposed model.

3. Application of the proposed model with different identifying constraints to examine how this impacts interpretation.

For the LMR example, the proposed longitudinal explanatory model is given by:

$$
\logit(\Pr(Y_{itp} = 1 \mid \theta_{tp})) = \theta_{tp} - \delta_{itp}
$$

$$
\theta_{pt} = \pi_{p0} + \pi_1\text{comp}_p + \pi_2\text{post}_p + \pi_3\text{final}_p \\
\quad + \pi_4\text{post}_p\text{comp}_p + \pi_5\text{final}_p\text{comp}_p \\
\pi_{p0} \sim N(0, \psi) \tag{3.19}
$$

$$
\delta_{itp} = \beta_{1tp}\text{fraction}_i + \beta_{2tp}\text{line}_i + \beta_{3tp}\text{CR}_i
$$

$$
\beta_{ktp} = \gamma_{k00} + \gamma_{k01}\text{comp}_p + \gamma_{k10}\text{post}_p + \gamma_{k20}\text{final}_p \\
\quad + \gamma_{k11}\text{post}_p\text{comp}_p + \gamma_{k21}\text{final}_p\text{comp}_p, \quad k = 1, 2, 3
$$

where \text{comp}_p is an indicator for being in the comparison group (versus LMR), \text{post}_p and \text{final}_p are indicators for being at posttest or end-of-year test (versus pretest), respectively, and \text{fraction}_i, \text{line}_i, and \text{CR}_i are indicators for the corresponding item features. To ease interpretation, the coefficients \gamma_{krs} have subscripts \(r = 0, 1, 2\) for covariates related to time and \(s = 0, 1\) for covariates related to person groups. Note that the model for \theta_{tp} differs from that given in Equation 3.4 because dummy variables are included for the occasions rather than assuming linear growth. As discussed on page 69, this model can be extended by including a random residual for items, \(\epsilon_i \sim N(0, \tau)\), in the top equation.
CHAPTER 3. LONGITUDINAL EXPLANATORY ITEM RESPONSE MODELS

Table 3.1: Description of the eight models fit to the LMR data.

<table>
<thead>
<tr>
<th>Model</th>
<th>Items</th>
<th>Explanatory?</th>
<th>Coding</th>
<th>Item Residual?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Common</td>
<td>Explanatory</td>
<td>Effect</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Common</td>
<td>Explanatory</td>
<td>Effect</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Common</td>
<td>Descriptive</td>
<td>Effect</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>All</td>
<td>Explanatory</td>
<td>Effect</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>All</td>
<td>Explanatory</td>
<td>Effect</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>All</td>
<td>Descriptive</td>
<td>Effect</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>Common</td>
<td>Explanatory</td>
<td>Dummy\textsuperscript{a}</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>Common</td>
<td>Explanatory</td>
<td>Dummy\textsuperscript{b}</td>
<td>No</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Reference items: integers, no line, multiple choice.
\textsuperscript{b} Reference items: integers, line, multiple choice.

The corresponding descriptive model is given by:

\[
\text{logit}(\Pr(Y_{itp} = 1 \mid \theta_{tp})) = \theta_{tp} - \delta_{itp}
\]

\[
\theta_{tp} = \pi_{p0} + \pi_{1}\text{comp}_p + \pi_{2}\text{post}_pt + \pi_{3}\text{final}_p
\]

\[
+ \pi_{4}\text{post}_pt\text{comp}_p + \pi_{5}\text{final}_pt\text{comp}_p
\]

\[
\pi_{p0} \sim N(0, \psi)
\]

\[
\delta_{itp} = \gamma_{i00} + \gamma_{i01}\text{comp}_p + \gamma_{i10}\text{post}_pt + \gamma_{i20}\text{final}_p
\]

\[
+ \gamma_{i11}\text{post}_pt\text{comp}_p + \gamma_{i21}\text{final}_pt\text{comp}_p
\]

To address the research issues, eight models were fit to the LMR data. Table 3.1 describes the eight models in terms of four factors: (1) whether the model was applied to only the common items or to all of the items, (2) whether the model was explanatory (Equation 3.19) or descriptive (Equation 3.20), (3) whether the model used effect or dummy coding, and (4) whether the model included a random item residual (where applicable). Models 1, 2, and 3, which included only the common items as an example of well-designed items, address the first research issue. Models 4, 5, and 6, which included the full set of items as an example of less well-designed items, address the second research issue. Models 7 and 8, which operationalized different identifying constraints using different model parameterizations, address the third research issue.

Estimation of all models was done using the \texttt{xtmelogit} command in Stata12 (StataCorp, 2011; for more detail see Rabe-Hesketh & Skrondal (2012, chapter 10)) except for estimation of the models with random item residuals which was done using the \texttt{lmer} command in R.
(R Development Core Team, 2012; for more detail see De Boeck, Bakker, Zwiter, Nivard, Hofman, Tuerlinckx & Partchev (2011)). `xtmelogit` uses adaptive quadrature with seven integration points by default while `lmer` uses the Laplace approximation. To check that switching between estimation routines does not alter the results, all models were fit using `lmer` in addition to Stata; the results were the same to the first decimal place. Since estimation using adaptive quadrature is expected to be reliable and less biased than estimation using the Laplace approximation for binary data (Rabe-Hesketh et al., 2005; Joe, 2008), adaptive quadrature results are reported when possible. All graphs were made using R.

3.3.3 Results

Application of the proposed longitudinal explanatory model to well-designed items

The proposed longitudinal explanatory model (Model 1, Equation 3.19) was fit to the LMR data for the common items with effect coding. The parameter estimates from fitting the longitudinal time-varying, differential item feature effects model to the LMR data are given in the first columns of Tables 3.2 and 3.3 (common items, fixed). Table 3.2 gives the estimated person regression coefficients. Table 3.3 gives the estimated item feature effect (i.e. the estimated difference in item difficulty between the relevant item groups), $\beta_{ktp}$, for each item feature at each occasion in each treatment group. For example, the first row gives $\beta_{100}$, the difference in item difficulty between fractions and integers items ($k = 1$) at pretest for students in the LMR group, which is estimated to be 1.58 logits (SE 0.09); the last row gives $\beta_{311}$, the difference in item difficulty between multiple choice (MC) and constructed response (CR) items at final test for students in the Comparison group, which is estimated to be 0.80 logits (SE 0.08). The results are shown graphically in Figure 3.3. The person side of Figure 3.3 displays the estimated average of $\theta_{tp}$ (using the estimates of $\pi_1$ to $\pi_5$ reported in the first column of Table 3.2) at each occasion in each group, while the item side displays $\hat{\delta}_{iptp}$ (calculated from the estimates in the first column of Table 3.3).
Table 3.2: Estimated person regression coefficients (and SEs) from Models 1 to 3 (common items) and 4 to 6 (all items); Models 1 and 4 are explanatory with only fixed item feature effects, Models 2 and 5 are explanatory with fixed item feature effects and a random item residual, and Models 3 and 6 are descriptive.

<table>
<thead>
<tr>
<th>Fixed Part</th>
<th>Common Items</th>
<th>All Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fixed</td>
<td>Residual</td>
</tr>
<tr>
<td>Comparison ($\pi_1$)</td>
<td>0.10 (0.12)</td>
<td>0.11 (0.13)</td>
</tr>
<tr>
<td>Post ($\pi_2$)</td>
<td>1.64 (0.06)</td>
<td>1.82 (0.07)</td>
</tr>
<tr>
<td>Final ($\pi_3$)</td>
<td>1.80 (0.06)</td>
<td>1.98 (0.07)</td>
</tr>
<tr>
<td>Comp×Post ($\pi_4$)</td>
<td>-1.32 (0.08)</td>
<td>-1.48 (0.09)</td>
</tr>
<tr>
<td>Comp×Final ($\pi_5$)</td>
<td>-0.53 (0.09)</td>
<td>-0.59 (0.09)</td>
</tr>
<tr>
<td>Random Part</td>
<td>var($\pi_{0p}$)</td>
<td>1.76 (0.12)</td>
</tr>
</tbody>
</table>

$^a \text{1mer}$ does not provide standard errors of random effects variances.
Table 3.3: Estimated item feature effects (and SEs) from Models 1 to 2 (common items) and Models 4 to 5 (all items); Models 1 and 4 are explanatory with only fixed item feature effects, and Models 2 and 5 are explanatory with fixed item feature effects and a random item residual.

<table>
<thead>
<tr>
<th>Item Feature</th>
<th>Test</th>
<th>Group</th>
<th>Parameters$^a$</th>
<th>Common Items</th>
<th>All Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Fixed</td>
<td>Residual</td>
</tr>
<tr>
<td>Fraction - Integer</td>
<td></td>
<td></td>
<td>$2(\gamma_{100} - \gamma_{101} - \gamma_{110} - \gamma_{120} + \gamma_{111} + \gamma_{121})$</td>
<td>1.58 (0.09)</td>
<td>1.66 (0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{100} + \gamma_{101} - \gamma_{110} - \gamma_{120} - \gamma_{111} - \gamma_{121})$</td>
<td>1.49 (0.09)</td>
<td>1.55 (0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{100} + \gamma_{101} - \gamma_{110} + \gamma_{111})$</td>
<td>0.79 (0.09)</td>
<td>0.77 (0.42)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{100} + \gamma_{101} - \gamma_{110} + \gamma_{111})$</td>
<td>1.80 (0.09)</td>
<td>1.84 (0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{100} + \gamma_{101} + \gamma_{120} - \gamma_{121})$</td>
<td>1.18 (0.10)</td>
<td>1.10 (0.42)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{100} + \gamma_{101} + \gamma_{120} + \gamma_{121})$</td>
<td>1.35 (0.09)</td>
<td>1.21 (0.41)</td>
</tr>
<tr>
<td>Line - No line</td>
<td></td>
<td></td>
<td>$2(\gamma_{200} - \gamma_{201} - \gamma_{210} - \gamma_{220} + \gamma_{211} + \gamma_{221})$</td>
<td>1.65 (0.09)</td>
<td>1.76 (0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{200} + \gamma_{201} - \gamma_{210} - \gamma_{220} - \gamma_{211} - \gamma_{221})$</td>
<td>1.65 (0.09)</td>
<td>1.76 (0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{200} + \gamma_{201} - \gamma_{210} + \gamma_{211})$</td>
<td>0.47 (0.09)</td>
<td>0.43 (0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{200} + \gamma_{201} + \gamma_{220} + \gamma_{221})$</td>
<td>1.46 (0.09)</td>
<td>1.48 (0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{200} - \gamma_{201} + \gamma_{220} - \gamma_{221})$</td>
<td>1.03 (0.09)</td>
<td>0.94 (0.41)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{200} + \gamma_{201} + \gamma_{220} + \gamma_{221})$</td>
<td>1.70 (0.09)</td>
<td>1.65 (0.41)</td>
</tr>
<tr>
<td>CR - MC</td>
<td></td>
<td></td>
<td>$2(\gamma_{300} - \gamma_{301} - \gamma_{310} - \gamma_{320} + \gamma_{311} + \gamma_{321})$</td>
<td>0.82 (0.08)</td>
<td>0.85 (0.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{300} + \gamma_{301} - \gamma_{310} - \gamma_{320} - \gamma_{311} - \gamma_{321})$</td>
<td>0.83 (0.08)</td>
<td>0.87 (0.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{300} + \gamma_{301} - \gamma_{310} + \gamma_{311})$</td>
<td>1.00 (0.09)</td>
<td>1.09 (0.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{300} + \gamma_{301} + \gamma_{320} + \gamma_{321})$</td>
<td>0.82 (0.08)</td>
<td>0.83 (0.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{300} - \gamma_{301} + \gamma_{320} - \gamma_{321})$</td>
<td>0.70 (0.09)</td>
<td>0.77 (0.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$2(\gamma_{300} + \gamma_{301} + \gamma_{320} + \gamma_{321})$</td>
<td>0.80 (0.08)</td>
<td>0.90 (0.39)</td>
</tr>
<tr>
<td>Residual Item Variance</td>
<td>var($\epsilon_i$)</td>
<td></td>
<td>0.55 (NA)$^b$</td>
<td>1.86 (NA)$^b$</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ The linear combination is multiplied by 2 due to the effect coding.

$^b$ lmer does not provide standard errors of random effects variances.
Looking at the growth of the LMR (L) and Comparison (C) students on the person side in Figure 3.3, we see that LMR students gained more in achievement than Comparison students during the year. At pretest, the estimated average proficiency was the same in the LMR and Comparison groups (difference of 0.1 logits, SE 0.1). The LMR students had higher estimated average growth than Comparison students from pre to post (1.3 logits, SE 0.1) and final (0.5 logits, SE 0.1). These results correspond to the findings from the efficacy study (Saxe et al., 2013), though the magnitude of the estimated effects on the logit scale are slightly smaller. Thus, the proposed model can be used to make a general evaluation of overall efficacy.

From Figure 3.3 and the first column of Table 3.3, we also see that the estimated effects of the item features differed between the LMR and Comparison groups. For Comparison students, there was little change in the effect of the item features over time, as seen in the mostly flat horizontal lines in the bottom panel of Figure 3.3. The only statistically significant effect was a slightly larger difference in difficulty between fractions and integers items at posttest. Since the instruction for the Comparison group was not designed around these item features, we would not expect the Comparison group to show changes in these item feature effects over time.

In contrast, for LMR students, there were a number of significant shifts in the effect of item features over time. The relative difficulty of fractions and number line items became substantially smaller at posttest, then increased again at final test. These were the two item features that were expected to be affected by the curriculum. The reduction in the difference in difficulty between fractions and integers content makes sense in light of the multiple connections drawn during the LMR curriculum between these domains. The reduction in the difference in difficulty between number line and non number line items is expected given the increased familiarity and emphasis on number line reasoning. The differential shifts in item feature effects over time thus reflect the underlying design of the LMR curriculum.

The descriptive model that corresponds to the proposed explanatory model was also estimated. Figure 3.4 and the third column of Table 3.2 (common items, descriptive) give the results from fitting the longitudinal time-varying, differential item difficulty model (Model 3, Equation 3.20) to the LMR data.

In general, we would compare the results in Figure 3.4 to Figure 3.3, and correspondingly the third and first columns of Table 3.2, to get a sense for how well the explanatory model fits the data. For example, the regression coefficients for person growth in the descriptive model are approximately 12% larger than in the explanatory model, better matching those from the original efficacy study. This suggests scale shrinkage in the explanatory model due to residual variance unexplained by the item features being absorbed by the (fixed) logistic error term. Standard model fit criteria, the Akaike and Bayesian information criteria (AIC, Akaike, 2002 and BIC, Schwarz, 1978), are given in Table 3.4. In the context of multilevel random effects models, it is not well established what values to use for the number of observations and the number of parameters when calculating AIC and BIC. We used the number of observations equal to the number of people and the number of parameters equal to the number of estimated parameters (e.g. the person abilities $\theta_{tp}$ were not counted as parameters). The model fit criteria also indicate that the descriptive model fits the data.
Figure 3.3: Results from Model 1: Time-varying, differential item feature effects for common items with effect coding.
Figure 3.4: Results from Model 3: Time-varying, differential item difficulties for common items with effect coding.
Table 3.4: Comparison of model fit between the explanatory and descriptive models for the common items.

<table>
<thead>
<tr>
<th>Model</th>
<th>$n_{pers}$</th>
<th>$n_{par}$</th>
<th>Log-likelihood</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 Explanatory</td>
<td>571</td>
<td>25</td>
<td>-13488.4</td>
<td>27026.8</td>
<td>27135.5</td>
</tr>
<tr>
<td>Model 2 Explanatory - Random</td>
<td>571</td>
<td>26</td>
<td>-12590.2</td>
<td>25232.4</td>
<td>25345.4</td>
</tr>
<tr>
<td>Model 3 Descriptive</td>
<td>571</td>
<td>103</td>
<td>-12340.9</td>
<td>24887.8</td>
<td>25335.5</td>
</tr>
</tbody>
</table>

better than the explanatory model.

Given the apparent scale shrinkage, the proposed explanatory model was modified to include a random residual for item. The results from applying this model (Model 2) to the LMR data are shown in the second column of Tables 3.3 and 3.2 (common items, residual). Once the random residual is added, the proposed explanatory model fits the LMR data as well as the descriptive model (Table 3.4). However, the confidence intervals overlap for the item feature effects and person growth estimates between Models 1 and 2. Therefore, the proposed explanatory model with or without the random residual fits the data well enough to support the interpretation of the estimated coefficients. This finding is based on typical model fit criteria (AIC and BIC), the use of which is not fully established for multilevel logistic random effects models.

In summary, when applying the proposed model to the well-designed common items, the model fits the data well enough to be used to evaluate the efficacy of the LMR curriculum. The proposed model separated changes in person achievement and item feature effects. The growth on the person side can be used to make a general evaluation of overall efficacy while the changes on the item side can be used to evaluate whether that growth was driven by the expected design features. In this case, the reduced difference in item difficulty between fractions and integers and between number line and non number line indicates that the LMR curriculum was effective due to the coherence provided by the single representational context.

**Application of the proposed model to less well-designed items**

To see how the use of the proposed model was affected if it was applied to less well-designed items, the proposed explanatory model was estimated using all of the LMR items (Model 4). The results are displayed graphically in Figure 3.5, with some of the corresponding estimates in the third column of Table 3.3 and the fourth column of Table 3.2 (all items, fixed).

Comparing the results from the model with all items (Model 4) to the results from the original model using only the common items (Model 1), we note two main differences. First, the estimates of both the item feature effects and the person growth coefficients from the model with all of the items are all closer to zero (i.e. appear “shrunk” towards zero) than the estimates from the model with only the common items. Second, the model with all of
Figure 3.5: Results from Model 4: Time-varying, differential item feature effects for all items with effect coding.
the items shows different patterns of person growth and item feature effects. For example, the Comparison students display negative growth from pretest to posttest, and there are statistically significant item feature effect changes over time.

There are a number of possible statistical explanations for these observed differences. The apparent shrinkage could perhaps be an effect of noise introduced because the parameters for the non-common items are being less precisely estimated than the parameters for the common items. Since the non-common items are seen on only one or two of the three occasions, they have one-third or two-thirds of the data as the common items. Examining the item difficulties from the descriptive model applied to all items from Saxe et al. (2013), we do observe that the standard errors of the item difficulty estimates are larger for the non-common items (mean 0.19, SE 0.002) than for the common items (mean 0.16, SE 0.02). However, the magnitude of this difference is small; with over 500 people, all of the item difficulties should be well estimated.

The apparent shrinkage could alternatively be due to attenuation of the whole latent scale when unexplained item variation is absorbed by the logistic error in the item response model, as discussed on page 69. When the estimated fixed item difficulties are regressed on the three primary item features, 52% of the variation in the difficulties of the common items is explained but only 19% of variation in the difficulties of all of the items is explained. This difference is not due to differences in the variance of the item features, as the item features are similarly distributed across the common items and the full set of items. This difference persists if the differential standard errors are accounted for in the regression (as in a random-effects meta-analysis regression). The poor explanatory power of the item features for the full set of items would lead to the observed attenuation.

To check this, a random residual for items was added to the model fit to all of the items; the results are given in the fourth column of Table 3.3 and the fifth column of Table 3.2 (all items, residual). The estimated variance of the random item residual is three times larger for all of the items than for the common items only. Once the random item residual is included, the person growth coefficients are no longer shrunk. The proposed model with a random item residual yields the same estimates for person growth for the common items and for all items (Table 3.2, columns 2 and 5); these models also yield the same estimates as the descriptive model (Table 3.2, columns 3 and 6). A large part of the observed differences between the model when applied to all of the items rather than just the common items was due to the difference in explained item difficulty.

Though the statistical modification of adding a random item residual leads to person growth coefficients that are no longer shrunk, the item feature effects for all items still show differences compared to the item feature effects for only the common items under the model with the random item residual (Table 3.3, column 4 compared to column 2). Thus, the effect of using less well-designed items for the proposed model cannot be fully corrected for by statistical means. The amount of difference differs among the item feature effects. This could occur if the less well-designed non-common items differed from the well-designed common item in two ways: (1) if the estimated item feature effects were smaller in magnitude; and (2) if the estimated item feature effects were different in relation to each other. An examination
1. Mark with an arrow (↑) where 47 belongs on the number line.

![Number Line](image)

2. Write the number that belongs in each box.

![Number Line](image)

3. Which mixed number equals $\frac{14}{6}$? Circle the answer.

A) $2 \frac{2}{14}$  
B) $2 \frac{1}{3}$  
C) $2 \frac{6}{14}$  
D) $3 \frac{1}{3}$

4. Write these fractions in order from least to greatest: $\frac{1}{5}$, $\frac{3}{10}$, $\frac{1}{4}$

Figure 3.6: Four example non-common items that are influential for estimating the item feature effects.

of the specific LMR items suggests that both of these occur.

Four of the most influential items from the regression of item difficulty on the item features are shown in Figure 3.6; they are all non-common items. The first two items share the same item feature classification (integer, line, constructed response), but represent one of the easiest and one of the hardest items. The difference in item difficulty within the same item feature cell due to important-but-not-modeled item features such as unit intervals versus unusual multiunit intervals would contribute to a smaller magnitude in the estimated feature effect of all features. The third and fourth items are each much harder than expected given one aspect of their item feature design: being multiple choice for the mixed number problem and being not on a number line for the ordering fractions problem. These both contribute to different relations between the item features than seen for the common items.

These results, from applying the proposed model to items where the modeled item features do not describe the differences in item difficulty well, suggest that how well the item features explain item difficulty is a key factor for the usefulness of the proposed model. There is no statistical difficulty in having non-common items (and therefore missing data), but as can be seen in this example, care must be taken in selecting the set of items to be
Table 3.5: Comparison of person regression estimates between effect and dummy coding.

<table>
<thead>
<tr>
<th></th>
<th>Effect Coding</th>
<th>Dummy Coding 1</th>
<th>Dummy Coding 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est. (SE)</td>
<td>Est. (SE)</td>
<td>Est. (SE)</td>
</tr>
<tr>
<td>Fixed Part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison ($\pi_1$)</td>
<td>0.1 (0.1)</td>
<td>0.1 (0.2)</td>
<td>0.1 (0.2)</td>
</tr>
<tr>
<td>Post ($\pi_2$)</td>
<td>1.6 (0.1)</td>
<td>0.7 (0.2)</td>
<td>1.9 (0.1)</td>
</tr>
<tr>
<td>Final ($\pi_3$)</td>
<td>1.8 (0.1)</td>
<td>1.2 (0.2)</td>
<td>1.8 (0.1)</td>
</tr>
<tr>
<td>Comp×Post ($\pi_4$)</td>
<td>-1.3 (0.1)</td>
<td>-0.4 (0.2)</td>
<td>-1.4 (0.2)</td>
</tr>
<tr>
<td>Comp×Final ($\pi_5$)</td>
<td>-0.5 (0.1)</td>
<td>-0.0 (0.2)</td>
<td>-0.7 (0.2)</td>
</tr>
<tr>
<td>Random Part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{var}(\pi_{0p})$</td>
<td>1.8 (0.1)</td>
<td>1.8 (0.1)</td>
<td>1.8 (0.1)</td>
</tr>
<tr>
<td>Model Statistics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log-Likelihood</td>
<td>-13488.4</td>
<td>-13488.4</td>
<td>-13488.4</td>
</tr>
<tr>
<td>$n_{par}$</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

used in the analysis. Though it is important to compare the explanatory model with and without a random item residual to ensure that there is no attenuation of the estimates, this statistical check and correction cannot fully rectify any problems caused by items that are contaminated by features that were not included as part of the item design.

In summary, it is important to carefully design the set of items to be used in this type of analysis around the desired item features and to then check whether the item feature design does adequately and accurately explain the item difficulties. The regression model for the item feature effects will generally be small (for the LMR data, there were only 41 unique items for three main effects). This increases the influence of each individual item on the results and makes it easier for a few “wild” items to alter the conclusions.

Application of the proposed model with different identifying constraints

To better understand the effect of the choice of identifying constraint on model interpretation, the proposed explanatory model was estimated using three different parameterizations: (1) effect coding of the items (Model 1); (2) dummy coding of the items with Integers, No Line, and Multiple Choice as the reference group (Model 7); and (3) dummy coding of the items with Integers, Line, and Multiple Choice as the reference group (Model 8). The results are displayed graphically in Figures 3.3, 3.7, and 3.8, respectively, with some of the corresponding estimates in Tables 3.5 and 3.6.

As discussed above, changing the parameterization does not change the statistical model being fit. This is reflected by the fact that Models 1, 7, and 8 all have the same log-likelihood and number of parameters (Table 3.5). Statistically, the models are equivalent.
Figure 3.7: Results from Model 7: Time-varying, differential item feature effects for common items with dummy coding (reference items: integers, no line, MC).
Figure 3.8: Results from Model 8: Time-varying, differential item feature effects for common items with dummy coding (reference items: integers, line, MC).
Table 3.6: Comparison of item feature effect estimates between effect and dummy coding.

<table>
<thead>
<tr>
<th>Item Feature</th>
<th>Occasion</th>
<th>Effect Coding</th>
<th>Dummy Coding 1</th>
<th>Dummy Coding 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LMR Comp LMR</td>
<td>LMR Comp LMR</td>
<td>LMR Comp LMR</td>
</tr>
<tr>
<td>Fraction - Integer</td>
<td>Pre</td>
<td>1.6 1.6 1.6</td>
<td>1.6 1.6 1.6</td>
<td>1.6 1.6 1.6</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>0.8 0.8 0.8</td>
<td>0.8 0.8 0.8</td>
<td>0.8 0.8 0.8</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>1.2 1.2 1.2</td>
<td>1.2 1.2 1.2</td>
<td>1.2 1.2 1.2</td>
</tr>
<tr>
<td>Line - No line</td>
<td>Pre</td>
<td>1.6 1.6 1.6</td>
<td>1.6 1.6 1.6</td>
<td>1.6 1.6 1.6</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>0.5 0.5 0.5</td>
<td>0.5 0.5 0.5</td>
<td>0.5 0.5 0.5</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>1.0 1.0 1.0</td>
<td>1.0 1.0 1.0</td>
<td>1.0 1.0 1.0</td>
</tr>
<tr>
<td>CR - MC</td>
<td>Pre</td>
<td>0.8 0.8 0.8</td>
<td>0.8 0.8 0.8</td>
<td>0.8 0.8 0.8</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>1.0 1.0 1.0</td>
<td>1.0 1.0 1.0</td>
<td>1.0 1.0 1.0</td>
</tr>
<tr>
<td></td>
<td>Final</td>
<td>0.7 0.7 0.7</td>
<td>0.7 0.7 0.7</td>
<td>0.7 0.7 0.7</td>
</tr>
</tbody>
</table>

Changing the parameterization also does not change the estimated item feature effects (Table 3.6). This is reflected in Figures 3.3, 3.7, and 3.8 by the fact that the order of and distance between the item estimates at each time point are the same. However, as can be seen in the figures, the estimated item difficulties change because the zero-point of the scale changes. For effect coding, the straight reference line is the average item difficulty; for the first dummy coding, it is the difficulty of the Integers, No line, MC items; for the second dummy coding, it is the Integers, Line, MC items. Thus, the relative movement of the item features looks different in the figures, but the estimated feature effects are unaffected.

However, changing the baseline via parameterization does alter the estimated person growth effects (see Table 3.5). For example, under the first dummy coding, the interaction term between Comparison group and final test is not statistically significant, indicating that the LMR students did not, on average, grow more than Comparison students between pre and final test. Under the second dummy coding, the estimated average proficiency of the LMR students declines between post and final tests. Both of these are in contrast to the findings from the effect coding parameterization as well as the more standard models in the efficacy study (Saxe et al., 2013). The parameterization does not appear to affect the estimate of the person residual variance.

By changing parameterization, and thereby changing what is being held constant in the model, we are changing the criterion that is being used to compare the person and item sides of the model. Under effect coding, person growth is relative to the average item difficulty of all of the items; under dummy coding, person growth is relative to the average item difficulty of
only those items being held constant. Thus, from the perspective of interpretation of finding of efficacy, the choice of model identification constraint matters. Because the average more closely reflects the traditional model, it is the more conservative choice.

3.4 Discussion

This paper proposed an extension to existing item response models that combined longitudinal growth on the person side with person- and time-varying item feature effects for the purpose of providing more nuanced looks at assessment-based efficacy results. From the application to the items that were well-described by the item features, it appears that the proposed model can be used to separate changes in person achievement and item effects.

There is evidence that the LMR curriculum was effective due to the single representational context and the link between the two content domains of integers and fractions, as hypothesized. In the LMR group, there was a reduced difference in the difficulty of items between fractions and integers and between number line and non number line, indicating that the overall growth was differentially driven by these factors. In contrast, there was almost no change in item feature effects over time for the Comparison group.

However, using items that are contaminated by item features other than the intended design features alters the results. Caution must be exercised in two ways: (1) to ensure that the item features adequately predict the item difficulties of the selected items and (2) to ensure that the selected items are a representative sample of the construct of interest. The affect of the first issue can be judged and partially (though not fully) accounted for by comparing an explanatory model with and without a random item residual in addition to comparing the explanatory model to a descriptive model. The affect of the second issue could possibly be explored by considering different subsets of items in a cross-validation-type procedure.

The analysis also revealed interesting features about the link between the model parameterization, the constraint used for identification, and the interpretation. In particular, it appears that effect coding, which implies that the identifying constraint is on the average of the item difficulties, is preferred from the point of view of interpretation because it gives results more consistent with the “standard” analysis. This parameterization reflects the more natural separation of person growth from item change. It also reflects a commonsense understanding of measurement invariance that can be used to interpret growth despite the invariant item parameters in the model.

One limitation of the current paper is the application to a single data set. Additional applications of the proposed model are needed to confirm its stability. Another limitation is the very simple growth model used on the person side. Further work could explore how interpretation changes if a more complex growth model, for example with additional random effects or multilevel effects, is used. A third possibility for further work would be to explore similar complications (such as interactions between the item features) on the item side.
Further work could also be done on model comparison criteria for the explanatory versus
descriptive models discussed herein.
Chapter 4

A longitudinal item response model for differential growth based on initial status

4.1 Introduction

There is a long-standing focus on equity in education in the United States. Outcomes and progress over time are compared between groups of students to examine issues of equity. The groups of interest have often been defined by student background characteristics (such as ethnicity, gender, or socioeconomic status) or student classification (such as English language status or eligibility for special education). Another important comparison is between students with different levels of prior knowledge. Examining how growth rates differ among students as a function of their current levels of achievement answers questions about how existing gaps in achievement change over time.

In addition to being a concern at the level of state and national policy, there is also a focus on equity in curriculum and instruction within classrooms. For example, in the Learning Mathematics through Representations (LMR) project, equity was one of the three main principles that guided the design of the new curriculum (Gearhart & Saxe, in press). The emphasis on differentiated instruction is reflected in the curriculum in multiple ways, such as tasks designed to be accessible to students with different levels of understanding and embedded opportunities for all students to contribute their ideas through partner work and class discussions. To check whether the curriculum supported growth for all students, one part of the overall evaluation of the LMR curriculum examined whether growth rates were similar for students with different levels of prior achievement (Saxe et al., 2013). Students were split into three groups based on their pretest score and growth rates were compared among these groups. Saxe et al. (2013) found that the mean growth rates were similar across all three groups, though students in the lower two groups grew slightly more than students in the highest group from pretest to final test. This analysis was a simple way to address the
research questions regarding whether students with different levels of prior knowledge were all learning the same amount under the LMR curriculum and, if not, whether students with a lower or higher level of prior knowledge were benefitting more.

These research questions regarding the relationship between prior knowledge and growth can be restated in more technical language:

1. Do students with different initial proficiency show differential change over time?
2. If so, is the relationship positive (such that the range of proficiency would increase from initial to final time) or negative (range decreases)?

To address these types of research questions in a more general way, this paper proposes longitudinal item response models for differential growth based on initial status. In the literature, this has been referred to more generally as the relationship between initial status and rate of change (Rogosa & Willett, 1985; Willett, 1988). Answering these questions is important for thinking about the consequences of educational programs at all levels. They address for whom those programs are more (or less) effective and whether the programs are expected to narrow or widen an existing achievement gap.

Statistical models for change over time were developed in parallel in three modeling traditions: hierarchical linear models (HLM), structural equation modeling (SEM), and item response modeling (IRM). Regression models for the relationship between initial status and change are found in both the HLM (e.g. Seltzer et al., 2003) and SEM (e.g. Muthén & Curran, 1997) literature; they are specific applications of the more general idea of latent variable regression (Raudenbush & Bryk, 2002, pp. 361-364). The models proposed in this paper extend previous HLM models to include a measurement model for a latent outcome variable; equivalently, the models extend previous longitudinal growth IRM models to include the regression of slope on initial status. Using latent variable models for the outcome rather than observed score models avoids the problem of regression to the mean.

In the next section, we review the literature on longitudinal growth modeling and previous discussions of the relationship between initial status and change. In Section 4.3, we propose a general longitudinal item response model for differential growth based on initial status and discuss the identification and estimation of the general model. The use and interpretation of the proposed model is illustrated with an empirical example in Section 4.4. The paper concludes with a discussion in Section 4.5.

4.2 Longitudinal growth modeling

First, we review models to analyze change over time. We integrate models from the traditions of hierarchical linear modeling, structural equation modeling, and item response modeling. Second, we review how previous models have addressed the relationship between initial status and rate of change.
4.2.1 Modeling change over time

First, we consider models for when the variable of interest is observed (or manifest). Second, we consider models for when the variable of interest is latent.

Outcome is manifest

Models for longitudinal data when the outcome of interest is observed were developed in both the HLM and SEM modeling frameworks.

Hierarchical linear modeling (HLM) The application of HLMs to longitudinal data considers repeated measurements as being clustered within individuals. In order to study change over time for individuals, we focus on a model for change within an individual, such as a model where the outcome at each time is regressed on a function of time rather than just a model for change in the population (Rogosa, Brandt & Zimowski, 1982). This leads to a two part representation for change containing both a model for individual growth and a model for individual differences in growth (Rogosa & Willett, 1985). These two parts fit naturally within an HLM framework, with a within-subject level 1 model for individual growth and a between-subject level 2 model for variation in growth between individuals (Goldstein, 1986; Bryk & Raudenbush, 1987).

A multilevel model for linear growth (Goldstein, 1986; Bryk & Raudenbush, 1987) is given by

\[
\begin{align*}
\text{Level 1} & \\
\text{Level 2} & \\
\end{align*}
\]

where \(y_{tp}\) is the observed measurement for person \(p\) at time \(t\) and \(\text{time}_{tp}\) gives the timing of the measurement occasions. \(\beta_{0p}\) is a person-specific intercept, often called the initial status; \(\gamma_{00}\) gives the mean at the initial time and \(u_{0p}\), called the random intercept, are person-specific deviations from that mean, assumed to follow a normal distribution with mean 0 and variance \(\psi_{00}\). \(\beta_{1p}\) is a person-specific rate of change, often called the rate of change or growth rate; \(\gamma_{10}\) is the mean rate of change and \(u_{1p}\), called the random slope, are person-specific deviations from that mean, assumed to follow a normal distribution with mean 0 and variance \(\psi_{10}\). \(\epsilon_{tp}\) are occasion-specific deviations from the person’s growth curve, assumed to follow a normal distribution with mean 0 and variance \(\tau\).
The linear growth model in Equations 4.1 to 4.5 is the simplest model for longitudinal data under the HLM framework. There are numerous extensions possible to this basic model. These include modeling more complicated growth curves such as polynomial or piecewise linear growth, including additional covariates at either or both levels to explain variation in growth, and including higher levels to account for additional clustering (Bryk & Raudenbush, 1987; Raudenbush, 1989). In order to fully capture the possibilities of longitudinal modeling using HLM, observations at more than two times are required (Rogosa & Willett, 1985; Bryk & Raudenbush, 1987). Raudenbush & Bryk (2002, chapter 6), Singer & Willett (2003, chapters 3-7), and Rabe-Hesketh & Skrondal (2012, chapter 7) contain comprehensive reviews of analyzing longitudinal data using multilevel models.

**Structural equation modeling (SEM)** The application of structural equation models to longitudinal data considers the covariance of repeated measurements over time. The same considerations that prompted the development of HLM models for longitudinal data, modeling both individual growth and a structure for combining individual growth functions, prompted SEM models for longitudinal data. McArdle & Epstein (1987) and Meredith & Tisak (1990) demonstrated how these goals could be accomplished within the framework of covariance structure analysis. They gave the name latent (growth) curve analysis to the procedure.

A structural equation model for growth is represented by the path diagram in Figure 4.1a. The model contains two latent variables (in circles), labeled “intercept” and “slope”, to account for the correlations between the observed measurements (in squares) at each time. These latent variables are called the growth factors; the intercept is often also called the initial status. The loadings from the intercept to each observed variable are fixed at 1. The loadings from the slope to each observed variable are fixed equal to the observation time (e.g. at 0, 1, 2, 3, resulting in a linear growth pattern); this requires the same timing of the measurement for each person. The means and covariance matrix for the intercept and slope latent variables are estimated. The residual variances of the manifest variables are also estimated, and are typically not constrained to be equal across occasions.

The latent curve model in Figure 4.1a is one of the simplest models for longitudinal data under the SEM framework. The general framework is very flexible. Possible extensions to this model include: adding time invariant and time-varying covariates (McArdle & Epstein, 1987; Willett & Sayer, 1994), adding additional latent variables for different growth structures (e.g. quadratic), estimating the slope loadings to estimate the shape of the growth curve (McArdle, 1988; Meredith & Tisak, 1990), using definition variables to allow for unequal timing of measurement (Mehta & West, 2000), and incorporating this growth process within a larger SEM model (Muthén & Curran, 1997). Willett & Sayer (1994), Muthén & Khoo (1998), Bollen & Curran (2006), and Preacher, Wichman, MacCallum & Briggs (2008) contain comprehensive reviews of analyzing longitudinal data using structural equation models.
(a) Correlated intercept and slope, observed measurement (Goldstein, 1986; Bryk & Raudenbush, 1987; McArdle & Epstein, 1987; Meredith & Tisak, 1990)

(b) Correlated intercept and slope, latent measurement (McArdle, 1988; Pastor & Beretvas, 2006; Wilson et al., 2012)

(c) Regression of slope on intercept, observed measurement (Muthén & Curran, 1997; Raudenbush & Bryk, 2002; Seltzer et al., 2003)

(d) Regression of slope on intercept, latent measurement

Figure 4.1: Path diagrams for four latent growth curve models.
CHAPTER 4. DIFFERENTIAL GROWTH ITEM RESPONSE MODEL

Connection between HLM and SEM  There is a “fundamental mathematical equivalence” between the models for longitudinal growth under the HLM and SEM frameworks (Willett & Sayer, 1994). For example, the models in Equations 4.1 to 4.5 and Figure 4.1a (with the slope loadings fixed for linear growth and residual variances constrained to be equal across time) are identical. The random intercept and coefficient in Equation 4.1 correspond to the two latent variables in Figure 4.1a. Willett & Sayer (1994), Singer & Willett (2003, chapter 8), and Rabe-Hesketh & Skrondal (2012, chapter 7, section 7) provide more detail on the correspondence between longitudinal data analysis using HLM and SEM.

Outcome is latent

Models for longitudinal data when the outcome of interest is latent were developed in the IRM framework in addition to both the HLM and SEM modeling frameworks.

Item response modeling (IRM)  In item response modeling, data are analyzed at the item level rather than after being aggregated to the person level. The observed variables are responses to questions. The variable for which we want to model change over time is a latent construct that is measured by the items. The item response model analyzes the relationship between the observed responses and the latent variable in terms of parameters representing the items and the people.

Models for analyzing longitudinal data within the IRM framework build upon the basic measurement model. Andersen (1985) and Embretson (1991) added additional latent variables to the person part of an item response model to capture change over time. Andersen (1985) included additional latent variables to represent the latent construct at each occasion while Embretson (1991) included additional latent variables to represent the change in the latent construct between successive occasions. Wilson et al. (2012) built upon these models, taking advantage of the flexible measurement model machinery afforded by the multidimensional random coefficients multinomial logit (MRCML) model framework (Adams et al., 1997), to specify latent growth item response models with specific functional forms for growth over time. Similarly, Pastor & Beretvas (2006), taking advantage of the connection between item response models and hierarchical generalized linear models (Kamata, 2001), presented a longitudinal item response model that allowed person proficiency and item difficulty to change over time.
An item response model for growth is given by:

Level 1
\[
\text{logit} \left[ \Pr(y_{itp} = 1 | \theta_{tp}) \right] = \theta_{tp} - \delta_i, \quad (4.6)
\]

Level 2
\[
\theta_{tp} = \beta_{0p} + \beta_{1p}\text{time}_{tp} + \epsilon_{tp}, \quad (4.7)
\]
\[
\epsilon_{tp} \sim N(0, \tau), \quad (4.8)
\]

Level 3
\[
\beta_{0p} = \gamma_{00} + u_{0p}, \quad (4.9)
\]
\[
\beta_{1p} = \gamma_{10} + u_{1p}, \quad (4.10)
\]
\[
\begin{pmatrix} u_{0p} \\ u_{1p} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{10} \\ \psi_{10} & \psi_{11} \end{pmatrix} \right), \quad (4.11)
\]

where \( y_{itp} \) is the observed item response to item \( i \) for person \( p \) at time \( t \) and \( \text{time}_{tp} \) gives the timing of the measurement occasions. \( \delta_i \) is the difficulty of item \( i \) and \( \theta_{tp} \) is the person- and time-specific latent proficiency. Notice that the Level 2 and 3 models in Equations 4.6 to 4.11 correspond exactly to the Level 1 and 2 models in Equations 4.1 to 4.5 with the observed \( y_{tp} \) replaced by the latent \( \theta_{tp} \). The interpretation of the remaining parameters is the same in both models. This model is represented by the path diagram in Figure 4.1c.

This model is similar to, though not identical to, the models of Wilson et al. (2012) and Pastor & Beretvas (2006). The latent growth item response model of Wilson et al. (2012) is more restrictive in some respects; it does not include the residuals at level 2 (i.e. leaves out the \( \epsilon_{tp} \) in Equation 4.7 and the small gray arrows going into the time-specific latent variables in Figure 4.1c). The models of Pastor & Beretvas (2006) and Wilson et al. (2012) are more general in other respects; they allow the item difficulty parameters \( \delta_i \) to vary over time by modeling differential item functioning (DIF) with respect to time.

**Connections to HLM and SEM** Models within the HLM and SEM frameworks have also been specified to model longitudinal data with latent variables of interest. In both, this involves incorporating a measurement model at the lowest level of the multilevel model (McArdle, 1988; Raudenbush, Rowan & Kang, 1991). In SEM, the model in Figure 4.1c has been called a curve of factors model (McArdle, 1988), a longitudinal model with multiple indicators (Muthén & Khoo, 1998), and a second-order latent growth model (Hancock, Kuo & Lawrence, 2001). This confluence of models is not surprising since all three modeling traditions fall under the larger statistical framework of generalized latent variable modeling (Bartholomew, 1987; Skrondal & Rabe-Hesketh, 2004).

**4.2.2 Relationship between initial status and change over time**

The relationship between initial status and change over time has been discussed in the statistical literature for decades. This discussion has focused primarily on estimating and
interpreting the correlation between initial status and rate of change with manifest outcome variables. The use of the term “initial” is based on the assumption that the occasion at which time is set equal to zero is the start of some process.

It will only make sense to discuss and model a specific correlation between initial status and change over time if two conditions are met. First, there needs to be meaningful variation in status at the time selected as the initial time. Second, there needs to be a substantive reason to define a specific time as the initial time (Willett, 1988). In educational research, there are a number of contexts, such as the commencement of a new program or new course of study, in which a specific initial time makes sense because the process being modeled has a defined specific starting point but there is also variation at that starting point as a result of previous processes.

The discussion of the correlation between initial status and growth relies on the existence of variation in status at the initial time. If the variance of one of two variables is zero, then the covariance between them is zero. If there is no variation in status at the initial time, then it does not make sense to discuss the correlation between initial status and rate of change. An example of this would be when initial status is zero for everyone at the selected initial time (i.e. when the initial time is a true start time), such as in vocabulary growth in young children (Huttenlocher, Haight, Bryk, Seltzer & Lyons, 1991). In these cases, the growth model may exclude a parameter for the variation in initial status (e.g. remove the random effect \(u_{0p}\) in Equation 4.9) though it still models variation in rate of change (i.e. retains the random effect \(u_{1p}\) in Equation 4.10).

Any discussion of a single correlation between initial status and rate of change also depends on which occasion is defined as the initial time (i.e. which occasion is labeled as time equal to zero). If we are not modeling a true start time when status is zero for everyone, the current status, even at the first measurement occasion, is a consequence of all previous experience. As a result, in longitudinal data, current status and growth are necessarily correlated, as noted by Rogosa et al. (1982) and Willett (1988). However, there is not a single true value for this correlation over time; the magnitude of the correlation between current status and growth depends on the measurement occasion (Rogosa & Willett, 1985; Willett, 1988). If there is a translation of the time scale, and therefore a different occasion is labeled as time equal to zero, the correlation between initial status (i.e. status when time is equal to zero) and growth will be different. As a result, any discussion of the correlation between initial status and change depends on which measurement occasion is selected as the initial time.

The estimation of the correlation between initial status and rate of change suffers from the problem of measurement error in an independent variable if the observed scores are used (Blomqvist, 1977). The correlation between the observed initial measurement and observed change is severely negatively biased (Blomqvist, 1977; Rogosa & Willett, 1985). This is an instance of the phenomenon of regression to the mean. As Rogosa et al. (1982) so eloquently puts it, “the correlation [between the true score and the rate of change] is an interesting fact of life, whereas the correlation [between the observed score and the difference score] is an uninteresting artifact of errors of measurement (p. 735)”. Various procedures have been
proposed to account for measurement error when estimating this correlation (e.g. Blomqvist, 1977; Rogosa & Willett, 1985).

However, when both the initial status and rate of change are modeled within an HLM, the true correlation between initial status and change is just the correlation between the person-specific slopes and intercepts (Bryk & Raudenbush, 1987). A consistent estimate can be obtained from the estimated covariance matrix of the random effects. Similarly, the correlation can be accurately estimated using the estimated covariance matrix of the latent variances in an SEM-based longitudinal model (Willett & Sayer, 1994).

To help interpret the correlation between initial status and rate of change, Figure 4.2 displays example linear growth trajectories with different correlations between initial status and rate of change. When the correlation is positive (Figure 4.2a), the slopes of the lines increase as initial status increases; this leads to increased variation and larger gaps over time. In contrast, when the correlation is negative (Figure 4.2b), the slopes of the lines decrease as initial status increases; this leads to reduced variation and smaller differences over time (though the gaps would increase again if the process continued further). There are two possibilities when the correlation is zero: (1) the slopes of the lines are constant (Figure 4.2c), resulting in parallel growth trajectories; or (2) the slopes of the lines vary in a manner unrelated to initial status (Figure 4.2d), leading to increased variation overall but no systematic increase or reduction in gaps.

The correlation provides an initial answer to the question about the relationship between initial status and change regarding the strength of the association. We may also be interested in the expected amount of change in the slope as initial status changes. To estimate this, we would regress the slope on the initial status. To do so, we could modify Equations 4.3 to 4.5 and Equations 4.9 to 4.11 as follows:

\[
\beta_{0p} = \gamma_{00} + u_{0p} \\
\beta_{1p} = \gamma_{10} + \gamma_{11} \beta_{0p} + u_{1p}
\]

\[
= \gamma_{10} + \gamma_{11}(\gamma_{00} + u_{0p}) + u_{1p}
\]

\[
\begin{pmatrix}
u_{0p} \\
u_{1p}
\end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix} , \begin{pmatrix} \psi_{00} & \psi_{10} \\ \psi_{10} & \psi_{11} \end{pmatrix} \right)
\]

\[
\psi_{10} = 0.
\]

This corresponds to modifying the path diagrams in Figures 4.1a and 4.1c to those in Figures 4.1b and 4.1d, respectively. Note that the correlation between the random effects is now fixed to 0. For manifest outcomes, this model has been discussed in both the HLM (Raudenbush & Bryk, 2002, chapter 11; Seltzer et al., 2003) and SEM (Muthén & Curran, 1997) literatures.

Switching between the correlation and the regression is simply a reparameterization of the model. The models represented by the top row and the bottom row of Figure 4.1 are equivalent and will fit empirical data equally well. A decision about which to use is specific to the research goals in a particular context.
In summary, to get an estimate of the expected difference in growth rates for people with different initial statuses, we can regress the slope of time on the initial status. The definition of the initial time (i.e. the selection of the zero point of the time scale or of which measurement occasion to label as time zero) needs to be purposeful in order for any interpretation to be sensible. We also need a variable for initial status that varies between individuals and is free from measurement error, which is straightforward to obtain in a model where initial status is represented by a latent variable. In a model where the measurement is manifest, this is the intercept term; in a model where the measurement is latent, there are more options, as will be discussed further in the next section.
4.3 A general item response model for differential growth

4.3.1 Model specification

A model for differential growth based on initial status can be used to address questions regarding different rates of change based on different starting points. In terms of the model, this means regressing the person-specific slope parameter $\beta_1$ on the latent variable representing the initial status. As can be seen in the different path diagrams in Figure 4.3, initial status could be defined in a few different ways in the statistical model.

For completeness, we start with a model in which the person-specific slope is not regressed
CHAPTER 4. DIFFERENTIAL GROWTH ITEM RESPONSE MODEL

107

on anything (Figure 4.3a); the equation for $\beta_{1p}$ is then (as in Equation 4.10):

$$\beta_{1p} = \gamma_{10} + u_{1p}.$$  \hfill (4.16)

As discussed above, this model should (as in the diagram) contain the correlation between the intercept and the slope, which gives some information about the relationship between growth and initial status.

Equivalently, we could regress the person-specific slope on the person-specific intercept (Figure 4.3b); in this case, the equation for $\beta_{1p}$ becomes (as in Equation 4.13):

$$\beta_{1p} = \gamma_{10} + \gamma_{11}\beta_{0p} + u_{1p} = \gamma_{10} + \gamma_{11}(\gamma_{00} + u_{0p}) + u_{1p}.$$  \hfill (4.17)

In this model, initial status is defined by the person-specific intercept $\beta_{0p}$. $\beta_{0p}$ draws information from all of the assessment occasions.

We could instead regress the person-specific slope on the person’s residual at the initial time (Figure 4.3c); in this case, the equation for $\beta_{1p}$ becomes:

$$\beta_{1p} = \gamma_{10} + \gamma_{12}\epsilon_{0p} + u_{1p}.$$  \hfill (4.18)

In this model, initial status is defined as the component of a person’s achievement at the initial time that is orthogonal to their general intercept. Regression on a residual instead of (or in addition to) regression on a latent variable such as the person-specific intercept has been used previously to distinguish between specific and general effects on an outcome variable (Bentler, 1990). The motivation and interpretation of regression on the residual is discussed further in the next section.

A third possibility would be to regress the person-specific slope on the person’s achievement at the initial time (Figure 4.3d); in this case, the equation for $\beta_{1p}$ becomes:

$$\beta_{1p} = \gamma_{10} + \gamma_{13}\theta_{0p} + u_{1p} = \gamma_{10} + \gamma_{13}(\gamma_{00} + u_{0p} + \epsilon_{0p}) + u_{1p}.$$  \hfill (4.19)

This model contains perhaps the most commonsense definition of initial status: achievement on the assessment at the initial time.

As is evident from comparing the equations, there is a general model for $\beta_{1p}$ that encompasses each of these options. The general form of the model is:

$$\beta_{1p} = \gamma_{10} + \gamma_{11}\gamma_{00} + \gamma_{12}\epsilon_{0p} + u_{1p}.$$  \hfill (4.20)

Each of the specific versions can be obtained by placing linear restrictions on parameters of the general form, as follows:

- $\gamma_{11} = \gamma_{12} = 0$: No regression (Equation 4.16)
Figure 4.4: Path diagram for the general model for differential growth based on initial status.

- $\gamma_{11} = 0$: Regress slope on intercept (Equation 4.17)
- $\gamma_{12} = 0$: Regress slope on residual at initial time (Equation 4.18)
- $\gamma_{11} = \gamma_{12} \equiv \gamma_{13}$: Regress slope on achievement at initial time (Equation 4.19)

If the general model is fit to data, we can use linear hypothesis tests of the regression coefficients to check the fit of the specific models.

The complete general item response model for differential growth, shown in Figure 4.4, is specified as follows:

**(Level 1)**

$$\text{logit}[Pr(y_{itp} = 1 \mid \theta_{tp})] = \theta_{tp} - \delta_i, \quad (4.21)$$

**(Level 2)**

$$\theta_{tp} = \beta_{0p} + \beta_{1p} \text{time}_{tp} + \epsilon_{tp} \quad (4.22)$$

$$\epsilon_{tp} \sim N(0, \tau), \quad (4.23)$$

**(Level 3)**

$$\beta_{0p} = \gamma_{00} + u_{0p} \quad (4.24)$$

$$\beta_{1p} = \gamma_{10} + \gamma_{11}(\gamma_{00} + u_{0p}) + \gamma_{12}\epsilon_{0p} + u_{1p} \quad (4.25)$$

$$\begin{pmatrix} u_{0p} \\ u_{1p} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{10} \\ \psi_{10} & \psi_{11} \end{pmatrix} \right) \quad (4.26)$$

$$\psi_{10} = 0. \quad (4.27)$$

Note that the arrow from time 0 to the slope is not needed in Figure 4.4 as it is effectively captured by the ones from the intercept and the time 0 residual; this is reflected by the final bullet point above.
In much of the following discussion of the model, we will focus on the higher levels of the model specifying the person growth. The combined form of the general model for $\theta_{tp}$ is:

$$
\theta_{tp} = (\gamma_{00} + u_{0p}) + \gamma_{10} \text{time}_{tp} + \gamma_{11}(\gamma_{00} + u_{0p})\text{time}_{tp} + \gamma_{12}\epsilon_{0p}\text{time}_{tp} + u_{1p}\text{time}_{tp} + \epsilon_{tp}
$$

$$
\epsilon_{tp} \sim N(0, \tau)
$$

$$
\begin{pmatrix}
    u_{0p} \\
    u_{1p}
\end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{10} \\ \psi_{10} & \psi_{11} \end{pmatrix} \right)
$$

$$
\psi_{10} = 0.
$$

4.3.2 Interpretation

The general model encompasses three options for defining “initial status”: the person-specific intercept ($\beta_{0p}$), the person’s residual at the initial time ($\epsilon_{0p}$), and the person’s achievement on the assessment at the initial time ($\theta_{0p}$). These three options are not independent of each other; the relationship between them is given by:

$$
\theta_{0p} = \beta_{0p} + \epsilon_{0p},
$$

$$
\text{corr}(\beta_{0p}, \epsilon_{0p}) = 0.
$$

The interpretation of the three options is intertwined as well. The person-specific intercept is the person’s expected achievement at the initial time, the residual is the deviation from expected achievement at the initial time, and the achievement is the actual achievement at the initial time.

The three options for defining initial status discussed here are analogous to the model components in latent state-trait theory (LST; Steyer, Ferring & Schmitt, 1992; Steyer, Schmitt & Eid, 1999). In LST, an observed score is decomposed into three components in two stages: first, the observed score is decomposed into the latent state (analogous to the actual achievement at the initial time) and measurement error; second, the latent state is decomposed into the latent trait (analogous to the expected achievement at the initial time) and the latent state residual (analogous to the deviation from expected achievement at the initial time). The latent trait characterizes the person and is defined as the expected score for a given person. The latent state characterizes the person in the situation and is defined as the expected score for a given person in a given situation, reflecting the interaction between person and situation. The latent state residual characterizes the measurement situation or occasion and is defined as the difference between the latent state and latent trait. Steyer et al. (1999) notes that the situation is defined by both observable or manipulable features of the measurement occasion and by unobservable psychological constructs, and that the latent state residual reflects both. The proposed model differs from the basic models in LST by (1) incorporating an item response model rather than modeling observed scores and (2) incorporating a growth structure for longitudinal observations to model the latent growth trajectory.
In the proposed model, the person-specific intercept corresponds to the person’s expected achievement at the initial time. In similar models, the intercept has been described as the true ability or status when time is equal to zero (Bryk & Raudenbush, 1987; Willett & Sayer, 1994). This statement makes sense when achievement is modeled as manifest (as in Figures 4.1a and 4.1b), because in these models, the intercept is free from measurement error while the manifest achievement at the initial time is not. However, when achievement is modeled as latent (as in the general model in Figure 4.4), both the intercept and the latent achievement at the initial time are free from measurement error, which is captured separately at the item level. In the general model, something other than measurement error distinguishes between the intercept and achievement at the initial time. The intercept must be interpreted more precisely as the expected achievement at the initial time.

The residual at the initial time corresponds to the deviation of actual achievement from expected achievement at the initial time. If achievement is modeled as manifest, the residual at the initial time confounds both measurement error and time-specific deviation. Once achievement is modeled as latent, and measurement error is captured separately at the item level, the residual represents only the time-specific deviation from the expected achievement. Since the linear growth model is only an approximation, as we always believe it to be, these deviations capture meaningful variation in achievement specific to the initial time. Further interpretation of these residuals depends on an understanding of the assessment context. The residuals capture both (1) deviations of the true growth trajectory from linearity and (2) observable and unobservable features of the initial assessment situation, such as material assessed only on the initial occasion, assessment conditions, and/or the influence of additional constructs such as person motivation that may fluctuate over time. Regressing the slope on the residual at the initial time in addition to the intercept allows us to simultaneously test the effect on growth of both the general expected achievement defined by the intercept and the unique aspects of achievement specific to the initial time.

The achievement at the initial time gives the most commonsense interpretation of initial status. If achievement is modeled as manifest, directly defining initial status as the manifest achievement is not advised unless adjustments are made for measurement error. Once achievement is modeled as latent, it becomes sensible to define initial status as the latent achievement. Initial status defined this way is interpreted as performance on the assessment at the initial time. Though this provides perhaps the most straightforward interpretation of initial status, it implies a statistical restriction that may not be justified: that the effect on growth is the same for both the intercept and the residual.

In the proposed general model, the effect on growth of the expected initial achievement and the deviation from that expectation are modeled simultaneously but separately. If the effects are similar, they can be interpreted as the effect of initial achievement. Otherwise, they should be interpreted separately.
4.3.3 Identification

The question of identification relates to whether there exists a unique set of estimates for the model parameters. A model is identified if there is a unique set of parameter values that generates the distribution of the observed data.

Demonstrating global identification is challenging. We can check for local identification, which is a necessary but not sufficient condition for global identification, using a method from a theorem by Rothenberg (1971). Rothenberg demonstrated that a model is locally identified for regular points if the rank of the Jacobian matrix of the reduced form parameters is equal to the number of fundamental parameters (as cited in Skrondal & Rabe-Hesketh, 2004, chapter 5). Intuitively, this procedure checks if there is a unique solution to the linear system of equations relating the parameters and the data distribution. Since the reduced form parameters are sufficient statistics for the data distribution, it suffices to check for the existence of a unique solution of the reduced form parameters in terms of the fundamental parameters.

Here, we focus on identification of the growth modeling part of the overall model, treating the latent $\theta_{tp}$ as observed. If there are more than two items at each time point, the measurement model part of the overall model should be identified. Because of assumptions of normality, all information on the distribution of $\theta_{tp}$ should be contained in first two moments.

To check the identification of the proposed general model, we first find the reduced form parameters by writing the first two moments of the observed data distribution in terms of the fundamental parameters:

$$E[\theta_{tp}] = (\gamma_{00} + (\gamma_{10} + \gamma_{11}\gamma_{00})time_{tp})$$
$$\text{var}[\theta_{tp}] = (\psi_{00} + \tau) + 2(\gamma_{11}\psi_{00} + \psi_{11} + \gamma_{12}\tau)time^{2}_{tp}$$
$$\text{cov}[\theta_{tp}, \theta_{t'p}] = (\psi_{00} + \tau)\text{time}_{tp}\text{time}_{t'p}$$

There are seven reduced form parameters to be solved for the seven fundamental parameters: $\gamma_{00}, \gamma_{10} + \gamma_{11}\gamma_{00}, \psi_{00}, \psi_{00} + \tau, \gamma_{11}\psi_{00}, \gamma_{11}\psi_{00} + \gamma_{12}\tau, \gamma_{11}\psi_{00} + \psi_{11} + \gamma_{12}\tau$. Note that if there are fewer than three time points observed, there are fewer than seven observed data equations to be solved for the seven parameters; thus, at least three time points are needed for the model to be identified.

Then, the Jacobian for the general model is given by:

$$\begin{bmatrix}
\gamma_{00} & 1 & 0 & 0 & 0 & 0 & 0 \\
\gamma_{10} + \gamma_{11}\gamma_{00} & \gamma_{11} & 1 & \gamma_{00} & 0 & 0 & 0 \\
\psi_{00} & 0 & 0 & 0 & 0 & 1 & 0 \\
\gamma_{11}\psi_{00} & 0 & 0 & \psi_{00} & 0 & 0 & \gamma_{11} \\
\psi_{00} + \tau & 0 & 0 & 0 & 1 & 1 & 0 \\
\gamma_{11}\psi_{00} + \gamma_{12}\tau & 0 & 0 & \psi_{00} + \tau & \gamma_{12} & \gamma_{11} & 0 \\
\gamma_{11}^{2}\psi_{00} + \psi_{11} + \gamma_{12}^{2}\tau & 0 & 0 & 2\gamma_{11}\psi_{00} + 2\gamma_{12}\tau & \gamma_{12}^{2} & \gamma_{11}^{2} & 1
\end{bmatrix}$$
Since the rank of this matrix is seven, the general model is locally identified for regular points.

To check the identification of the previously discussed restricted versions of the general model, we consider whether the rank of the Jacobian changes when the restricted parameter values are substituted into the matrix. When $\gamma_{11} = 0$, $\gamma_{12} = 0$, or $\gamma_{11} = \gamma_{12}$, the rank of the Jacobian remains seven. Thus, these are regular points, the model is locally identified at these points, and the restricted models are also locally identified.

However, the model is not locally identified for $\tau = 0$ or $\psi_{00} = 0$. When $\tau = 0$, the column corresponding to $\gamma_{12}$ becomes a column of all zeros; when $\psi_{00} = 0$, the columns corresponding to $\gamma_{10}$ and $\gamma_{11}$ each have an entry in only the second row. The rank of the Jacobian changes to six at either of these points, demonstrating that they are not regular points. These “irregular” points correspond to situations regressing the slope on a (latent) covariate with no variance, which has no unique solution. Convergence problems may arise if this model is fit to data where $\tau$ or $\psi$ is close to zero.

### 4.3.4 Equivalent models

There are a number of reparameterizations that are equivalent to the general model given in Equations 4.21 to 4.27. These equivalent models, shown in the path diagrams in Figure 4.5, arise by replacing either or both of the regressions of the slope with correlations (Figures 4.5b to 4.5d) or by setting the regression coefficients equal (i.e. regression on the initial achievement) and including a correlation between the slope and either the intercept or the initial residual (Figures 4.5e and 4.5f). These reparameterizations are useful because different points are regular (or not regular) under the different parameterizations. We illustrate this by showing the equivalence and identification for two of the alternative parameterizations.

The equation for $\theta_{tp}$ in the model with regression of the slope on the initial residual and with correlation between intercept and slope (Figure 4.5b) is given by:

$$
\theta_{tp} = (\gamma_{00}^* + u_{0p}) + \gamma_{10}^* \text{time}_{tp} + \gamma_{12}^* \epsilon_{0p} \text{time}_{tp} + u_{1p} \text{time}_{tp} + \epsilon_{tp}
$$

$$
\epsilon_{tp} \sim N(0, \tau^*)
$$

$$
\begin{pmatrix}
u_{0p} \\
u_{1p}
\end{pmatrix} \sim N
\begin{pmatrix}
0 \\
0
\end{pmatrix}
, \begin{pmatrix}
\psi_{00}^* & \psi_{10}^* \\
\psi_{10}^* & \psi_{11}^*
\end{pmatrix}
$$

This model has seven fundamental parameters: $\gamma_{00}^*$, $\gamma_{10}^*$, $\gamma_{12}^*$, $\tau^*$, $\psi_{00}^*$, $\psi_{11}^*$, and $\psi_{10}^*$. The observed data distribution is a function of these parameters as follows:

$$
E[\theta_{tp}] = (\gamma_{00}^*) + (\gamma_{10}^*) \text{time}_{tp}
$$

$$
\text{var}[\theta_{tp}] = (\psi_{00}^* + \tau^*) + 2(\psi_{10}^*) \text{time}_{tp} + (\psi_{11}^* + \gamma_{12}^* \tau^*) \text{time}_{tp}^2
$$

$$
\text{cov}[\theta_{0p}, \theta_{tp}] = (\psi_{00}^*) + (\psi_{10}^* + \gamma_{12}^* \tau^*) \text{time}_{tp}
$$

$$
\text{cov}[\theta_{tp}, \theta_{tp'}] = (\psi_{00}^*) + (\psi_{10}^*) \text{time}_{tp} + \text{time}_{tp'} + (\psi_{11}^* + \gamma_{12}^* \tau^*) \text{time}_{tp} \text{time}_{tp'}
$$
(a) General model

(b) Correlation between slope and intercept

(c) Correlation between slope and residual

(d) Both correlations

(e) Regress on achievement, correlation between slope and intercept

(f) Regress on achievement, correlation between slope and residual

Figure 4.5: Path diagrams for six models that are equivalent.
CHAPTER 4. DIFFERENTIAL GROWTH ITEM RESPONSE MODEL

This model has seven reduced form parameters, each of which corresponds to (i.e. enters into these equations in the same way as) one of the reduced form parameters from the original parameterization of the general model: \( \gamma_{00}^*, \gamma_{10}^*, \psi_{00}^*, \psi_{00}^* + \tau^*, \psi_{10}^*, \psi_{10}^* + \gamma_{12}^* \tau^*, \psi_{11}^* + \gamma_{12}^* \tau^* \).

The equation for \( \theta_{tp} \) in the model with regression of the person-specific slope on the initial achievement and with correlation between the slope and intercept (Figure 4.5e) is given by:

\[
\theta_{tp} = (\gamma_{00}^* + u_{tp}) + \gamma_{10}^* \text{time}_{tp} + \gamma_{11}^* (\gamma_{00}^* + u_{tp} + \epsilon_{tp}) \text{time}_{tp} + u_{1p} \text{time}_{tp} + \epsilon_{tp}
\]

\[
\epsilon_{tp} \sim N(0, \tau')
\]

\[
(\begin{pmatrix} u_{0p} \\ u_{1p} \end{pmatrix}) \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{10} \\ \psi_{10} & \psi_{11} \end{pmatrix}).
\]

This model has seven fundamental parameters: \( \gamma_{00}^*, \gamma_{10}^*, \gamma_{11}^*, \tau^*, \psi_{00}^*, \psi_{11}^* \), and \( \psi_{10}^* \). The observed data distribution is a function of these parameters as follows:

\[
E[\theta_{tp}] = (\gamma_{00}^*) + (\gamma_{10}^* + \gamma_{11}^* \gamma_{00}^*) \text{time}_{tp}
\]

\[
\text{var}[\theta_{tp}] = (\psi_{00}^* + \tau^*) + 2(\gamma_{11}^* \psi_{00}^* + \psi_{10}^*) \text{time}_{tp} + (\gamma_{11}^* \gamma_{00}^* + \psi_{11}^* + 2\gamma_{11}^* \psi_{10}^* + \gamma_{12}^* \tau^*) \text{time}_{tp}^2
\]

\[
\text{cov}[\theta_{0p}, \theta_{1p}] = (\psi_{00}^*) + (\gamma_{11}^* \psi_{00}^* + \psi_{10}^* + \gamma_{11}^* \tau^*) \text{time}_{tp} \text{time}_{tp}^p
\]

\[+ (\gamma_{11}^* \gamma_{00}^* + \psi_{11}^* + 2\gamma_{11}^* \psi_{10}^* + \gamma_{12}^* \tau^*) \text{time}_{tp}^p \text{time}_{tp}^p
\]

This model has seven reduced form parameters, which again correspond the reduced form parameters from the other parameterizations: \( \gamma_{00}^* \), \( \gamma_{10}^* + \gamma_{11}^* \gamma_{00}^* \), \( \psi_{00}^* \), \( \psi_{00}^* + \tau^* \), \( \gamma_{11}^* \gamma_{00}^* + \psi_{10}^* \), \( \gamma_{11}^* \gamma_{00}^* + \psi_{10}^* + \gamma_{11}^* \tau^* \), \( \gamma_{11}^* \gamma_{00}^* + \psi_{11}^* + 2\gamma_{11}^* \psi_{10}^* + \gamma_{12}^* \tau^* \).

We demonstrate the equivalence between these models and the proposed general model by showing that there is a closed form solution for the parameters of each of these models in terms of the parameters of the general model. We need only solve the following equations setting the corresponding reduced form parameters equal:

\[
\begin{align*}
\gamma_{00}^* &= \gamma_{00} \\
\gamma_{10} + \gamma_{11} \gamma_{00}^* &= \gamma_{10}^* + \gamma_{11}^* \gamma_{00} \\
\psi_{00}^* &= \psi_{00} \\
\psi_{00} + \tau^* &= \psi_{00} + \tau^* \\
\gamma_{11} \psi_{00}^* &= \gamma_{11}^* \psi_{00}^* + \psi_{10}^* \\
\gamma_{11} \psi_{00}^* + \gamma_{12} \tau^* &= \gamma_{11}^* \psi_{00}^* + \psi_{10}^* + \gamma_{11}^* \tau^* \\
\gamma_{11}^2 \psi_{00}^* + \gamma_{12} \tau^* &= \gamma_{11}^2 \psi_{00}^* + \psi_{11}^* + 2\gamma_{11}^* \psi_{10}^* + \gamma_{12}^* \tau^* = \psi_{11}^* + \gamma_{12}^* \tau^* \\
\end{align*}
\]

Solving these equations for the parameters of the two alternatives in terms of the proposed
general model, we obtain:

\[
\begin{align*}
\gamma_{00}^* &= \gamma_{00} \\
\gamma_{10}^* &= \gamma_{10} + \gamma_{00}\gamma_{11} \\
\gamma_{12}^* &= \gamma_{12} \\
\tau^* &= \tau \\
\psi_{00}^* &= \psi_{00} \\
\psi_{11}^* &= \psi_{11} + \psi_{00}\gamma_{11}^2 \\
\psi_{10}^* &= \psi_{00}\gamma_{11}
\end{align*}
\]

and

\[
\begin{align*}
\gamma_{00}' &= \gamma_{00} \\
\gamma_{10}' &= \gamma_{10} + \gamma_{00}(\gamma_{11} - \gamma_{12}) \\
\gamma_{11}' &= \gamma_{12} \\
\tau' &= \tau \\
\psi_{00}' &= \psi_{00} \\
\psi_{11}' &= \psi_{11} + \psi_{00}(\gamma_{11} - \gamma_{12})^2 \\
\psi_{10}' &= \psi_{00}(\gamma_{11} - \gamma_{12})
\end{align*}
\]

Therefore, the three parameterizations are equivalent.

Like the general model, these equivalent parameterizations are also locally identified for regular points. The Jacobian for the model in Figure 4.5b is

\[
\begin{bmatrix}
\gamma_{00}^* & \gamma_{10}^* & \gamma_{12}^* & \tau^* & \psi_{00}^* & \psi_{11}^* & \psi_{10}^* \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 2\gamma_{121}^* & 0 & 0 & 1 \\
0 & 0 & 2\gamma_{121}^* & \gamma_{12}^* & 0 & 0 & 1 \\
\end{bmatrix}
\]

and the Jacobian for the model in Figure 4.5e is

\[
\begin{bmatrix}
\gamma_{00} & \gamma_{10} & \gamma_{11} & \tau' & \psi_{00} & \psi_{11} & \psi_{10} \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\gamma_{11}' & 1 & \gamma_{00}' & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & \psi_{00}' & 0 & \gamma_{11}' & 0 & 1 \\
0 & 0 & \psi_{00}' & \gamma_{11}' & 0 & 1 & 0 \\
0 & 0 & 2\gamma_{111}(\psi_{00}' + \tau') + 2\psi_{10}' & \gamma_{11}' & \gamma_{11}' & 0 & 1 \\
0 & 0 & 2\gamma_{111}(\psi_{00}' + \tau') + 2\psi_{10}' & \gamma_{11}' & \gamma_{11}' & 1 & 2\gamma_{11}' \\
\end{bmatrix}
\]
The rank of each Jacobian is again seven, demonstrating local identification for regular points.

As for the general model, if $\tau = 0$, the rank of both of these Jacobians becomes 6, so this is not a regular point and the model is not locally identified under any of these parameterizations. However, if $\psi_{00} = 0$, the rank of these Jacobians does not change. In these parameterizations, this is a regular point and the model is locally identified. In these parameterizations the slope is no longer regressed on the intercept; instead, the correlation (or covariance) is modeled. So, a variance of zero for the intercept implies a covariance of zero rather than implying an infinite or indeterminate parameter value. Similarly, the parameterizations with the correlation between the slope and the initial residual (as in Figures 4.5c, 4.5d, and 4.5f) should be locally identified when $\tau = 0$.

If model estimation using one parameterization encounters convergence problems due to one of these irregular points, an alternative parameterization can be fit to the data instead. Because there are closed form solutions for calculating the parameters in one parameterization using the parameters from another parameterization, estimates of the parameters under the originally desired parameterization can then be calculated.

4.3.5 Estimation

Due to the presence in the general model of non-standard regressions of a latent variable (the slope) on a residual (the residual at the initial time), standard software needs to be “tricked” into estimating the model. We discuss how the general model can be estimated using the \texttt{gllamm} command (Rabe-Hesketh et al., 2004b) in Stata (StataCorp, 2011) and using Mplus (Muthén & Muthén, 2010). For both software packages, you need to define the residual at the initial time, $\epsilon_{0p}$, as a separate latent variable with appropriate restrictions.

Using \texttt{gllamm} in Stata

The model being estimated is a three-level model: level 1 is item, level 2 is time, and level 3 is person. Consider the combined level 2 and level 3 model for $\theta_{tp}$ given by Equation 4.28:

$$
\theta_{tp} = (\gamma_{00} + u_{0p}) + \gamma_{10}\text{time}_{tp} + \gamma_{11}(\gamma_{00} + u_{0p})\text{time}_{tp} + \gamma_{12}\epsilon_{0p}\text{time}_{tp} + u_{1p}\text{time}_{tp} + \epsilon_{tp} \quad (4.34)
$$

Note that $\epsilon_{0p}$ affects the responses for person $p$ at all occasions via the interaction term $\gamma_{12}\epsilon_{0p}\text{time}_{tp}$. We therefore need to model it as a separate latent variable at level 3 and constrain its variance to be equal to that of $\epsilon_{tp}$ for $t > 0$ at level 2.

The model therefore has three latent variables at level 3 ($\xi_{1p}^{(3)} \equiv \epsilon_{0p}$, $\xi_{2p}^{(3)} \equiv u_{0p}$, and $\xi_{3p}^{(3)} \equiv u_{1p}$) and one latent variable at level 2 ($\xi_{1tp}^{(2)} \equiv \epsilon_{tp}$ for $t > 0$). We also define $\eta_{2p}^{(3)} \equiv \gamma_{00} + \xi_{2p}^{(3)}$. 
Substituting these into Equation 4.34, we get:

\[
\theta_{tp} = \eta_{2p}^{(3)} + \gamma_{10} \text{time}_{tp} + \gamma_{11} \eta_{2p}^{(3)} \text{time}_{tp} + \gamma_{12} \xi_{1p}^{(3)} \text{time}_{tp} + \zeta_{3p}^{(3)} \text{time}_{tp}
\]

\[
+ \zeta_{1p}^{(3)} I(t = 0) + \zeta_{11p}^{(2)} I(t > 0)
\]

\[
= \gamma_{10} \text{time}_{tp} + \text{time}_{tp} \xi_{3p}^{(3)} + (1 + \gamma_{11} \text{time}_{tp}) \eta_{2p}^{(3)} + (I(t = 0) + \gamma_{12} \text{time}_{tp}) \xi_{1p}^{(3)}
\]

\[
+ I(t > 0) \zeta_{11p}^{(2)}
\]

This equation has 5 terms. The first term gives the fixed regression effect of time. The next four terms contain the four latent variables and linear predictors multiplying the latent variables. The linear predictors are treated by gllamm as equations for factor loadings for the latent variables (Rabe-Hesketh et al., 2004a). In each, the first loading is set to 1 and the remaining loadings are estimated; only the equations for the linear predictors for \( \eta_{2p}^{(3)} \) and \( \zeta_{3p}^{(3)} \) have more than one term, so only \( \gamma_{11} \) and \( \gamma_{12} \) are estimated. An unstructured covariance matrix is estimated for \( \zeta_{3p}^{(3)} \) and \( \eta_{2p}^{(3)} \). \( \xi_{1p}^{(3)} \) and \( \zeta_{11p}^{(2)} \) are constrained to have the same variance and to be uncorrelated with each other or the other latent variables.

An example of the gllamm syntax used to fit the general model to the data in the empirical application is given in Appendix 4.A.1.

### Using Mplus

Consider the path diagram in Figure 4.6. This path diagram gives the same model as the path diagram for the general model in Figure 4.4. The main difference between the path diagrams is the inclusion of an additional latent variable, labeled “Resid”.

In order to estimate the proposed general model in Mplus, we need to define an additional latent variable to represent the residual at the initial time. This latent variable has only a
single indicator, the latent variable at the initial time, whose loading is fixed to 1. The variance of the latent variable representing the residual at the initial time is constrained to be equal to the variances for the latent variables at the other times. The residual latent variable is further constrained to be uncorrelated with any of the other latent variables.

The other components of the model are defined in a standard way as follows. A latent variable is defined for each time point with the items at that time point as indicators. The thresholds (i.e. item difficulties) are constrained to be the same for each factor (i.e. to be the same over time) and all factor loadings are fixed to 1. The variances are constrained to the same and the latent variables are constrained to be uncorrelated with each other or with the other latent variables. A linear growth model is defined for the latent variable at each time point using additional latent variables for the intercept and slope; the factor loadings for the intercept are all fixed at 1 and the factor loadings for the slope are fixed to a linear growth pattern (i.e. 0, 1, 2, 3), which requires equal timing for all individuals. Finally, the latent variable for the slope is regressed on the latent variables for the intercept and for the residual at the initial time.

An example of the Mplus syntax used to fit the general model to the data in the empirical application is given in Appendix 4.A.2.

Parameter recovery

Given the complexity of the general model and in particular the non-standard regression of a latent variable on a residual, a simulation was conducted to examine the recovery of the parameters for the general model. The design for the simulation was taken from the empirical illustration in Section 4.4, which has 18 items given at 4 occasions. The original intention was to use a population study, wherein the response probabilities are computed for all possible response patterns and then used as frequency weights in the estimation (cf. Rotnitzky & Wypij, 1994; Heagerty & Kurland, 2001; Jeon, 2012). However, this was not feasible, as 18 items and 4 occasions results in $2^{18\times4} = 4.7 \times 10^{21}$ response patterns. Even reducing the size to 6 items and 3 occasions would result in an unmanageable $2^{6\times3} = 262144$ response patterns. Therefore, the desired population study was approximated by using finite samples with 5000 and 10000 people (simulated independently).

Data were simulated using the proposed general model (Equations 4.21 to 4.27) with parameters fixed to the values from the empirical data. The generating values are given in Table 4.1. The general model was then fit to the simulated data using Mplus with adaptive rectangular (i.e. Cartesian) quadrature with 10 integration points. Mplus was used rather than gllamm due to greatly reduced computation time; as will be seen in the empirical results in the next section, results are similar between Mplus and gllamm. Model estimation for the simulation took 5 hours with 5000 people and 20 hours with 10000 people.

The results from the simulation are given in Table 4.1. With 5000 people, all but one of the twenty-one fixed effect estimates (item difficulties and growth coefficients) were within twice their estimated standard error from the generating value (Figure 4.7a). With 10000 people, three of the item difficulties were more than twice their estimated standard error away
Table 4.1: Results from simulations for parameter recovery.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Gen. Value</th>
<th>Estimate (SE)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N = 5000</td>
<td>N = 10000</td>
<td></td>
</tr>
<tr>
<td><strong>Item Difficulties</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 1 ($\delta_1$)</td>
<td>-1.138</td>
<td>-1.073 (0.027)</td>
<td>-1.116 (0.019)</td>
<td></td>
</tr>
<tr>
<td>Item 2 ($\delta_2$)</td>
<td>-0.446</td>
<td>-0.425 (0.025)</td>
<td>-0.432 (0.018)</td>
<td></td>
</tr>
<tr>
<td>Item 3 ($\delta_3$)</td>
<td>0.679</td>
<td>0.701 (0.024)</td>
<td>0.687 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 4 ($\delta_4$)</td>
<td>0.794</td>
<td>0.841 (0.024)</td>
<td>0.817 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 5 ($\delta_5$)</td>
<td>1.579</td>
<td>1.610 (0.024)</td>
<td>1.613 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 6 ($\delta_6$)</td>
<td>0.198</td>
<td>0.207 (0.024)</td>
<td>0.224 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 7 ($\delta_7$)</td>
<td>0.329</td>
<td>0.326 (0.024)</td>
<td>0.341 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 8 ($\delta_8$)</td>
<td>-0.418</td>
<td>-0.377 (0.025)</td>
<td>-0.405 (0.018)</td>
<td></td>
</tr>
<tr>
<td>Item 9 ($\delta_9$)</td>
<td>-0.958</td>
<td>-0.933 (0.026)</td>
<td>-0.949 (0.019)</td>
<td></td>
</tr>
<tr>
<td>Item 10 ($\delta_{10}$)</td>
<td>-0.197</td>
<td>-0.166 (0.025)</td>
<td>-0.187 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 11 ($\delta_{11}$)</td>
<td>2.309</td>
<td>2.334 (0.026)</td>
<td>2.350 (0.018)</td>
<td></td>
</tr>
<tr>
<td>Item 12 ($\delta_{12}$)</td>
<td>-0.901</td>
<td>-0.889 (0.026)</td>
<td>-0.878 (0.018)</td>
<td></td>
</tr>
<tr>
<td>Item 13 ($\delta_{13}$)</td>
<td>1.667</td>
<td>1.685 (0.025)</td>
<td>1.691 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 14 ($\delta_{14}$)</td>
<td>-1.560</td>
<td>-1.570 (0.029)</td>
<td>-1.556 (0.020)</td>
<td></td>
</tr>
<tr>
<td>Item 15 ($\delta_{15}$)</td>
<td>0.125</td>
<td>0.111 (0.024)</td>
<td>0.157 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 16 ($\delta_{16}$)</td>
<td>0.552</td>
<td>0.549 (0.024)</td>
<td>0.554 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 17 ($\delta_{17}$)</td>
<td>0.834</td>
<td>0.817 (0.024)</td>
<td>0.860 (0.017)</td>
<td></td>
</tr>
<tr>
<td>Item 18 ($\delta_{18}$)</td>
<td>1.673</td>
<td>1.685 (0.025)</td>
<td>1.712 (0.017)</td>
<td></td>
</tr>
<tr>
<td><strong>Growth Coefficients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope ($\gamma_{10}$)</td>
<td>0.782</td>
<td>0.778 (0.006)</td>
<td>0.784 (0.005)</td>
<td></td>
</tr>
<tr>
<td>Slope on Inter ($\gamma_{11}$)</td>
<td>0.089</td>
<td>0.086 (0.009)</td>
<td>0.100 (0.006)</td>
<td></td>
</tr>
<tr>
<td>Slope on Resid ($\gamma_{12}$)</td>
<td>1.076</td>
<td>1.029 (0.092)</td>
<td>0.990 (0.060)</td>
<td></td>
</tr>
<tr>
<td><strong>Growth Variances</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ($\psi_{00}$)</td>
<td>1.125</td>
<td>1.137 (0.031)</td>
<td>1.105 (0.022)</td>
<td></td>
</tr>
<tr>
<td>Slope ($\psi_{11}$)</td>
<td>0.030</td>
<td>0.032 (0.010)</td>
<td>0.039 (0.007)</td>
<td></td>
</tr>
<tr>
<td>Residual ($\tau$)</td>
<td>0.103</td>
<td>0.099 (0.007)</td>
<td>0.109 (0.005)</td>
<td></td>
</tr>
</tbody>
</table>

from the generating value (Figure 4.7b). These rates are tolerable. The growth coefficient for the slope regressed on the intercept is the least precisely estimated parameter in both simulations, as reflected by the much larger standard error. The variance components were well-recovered with both numbers of people.
4.4 Empirical Application

The purpose of this section is to demonstrate the use of the proposed model for differential growth. Data from the implementation of a new mathematics curriculum for elementary school is used to illustrate the interpretation of the different specifications for initial status and the comparison between them.

4.4.1 Data

The Learning Mathematics through Representations (lmr) project has developed a supplemental curriculum for integers and fractions in the late-elementary grades. The curriculum consists of two two-week sequences of instruction, one for integers (9 lessons) and one for fractions (10 lessons). Two central features of the curriculum are (1) the use of the number line as a central representation and (2) the use of a series of principles and definitions that reflect the core ideas of the curriculum and are explicitly discussed and agreed upon by the students.

An evaluation of the curriculum was conducted in 2010-2011 in 11 treatment and 10 comparison classrooms in the San Francisco Bay Area (Saxe et al., 2013). Part of this study examined whether growth rates were similar for students with different levels of prior achievement. To explore this, the students were split into three groups based on their pretest
score, and growth rates were compared among these groups. Saxe et al. (2013) found that the growth rates were similar across all three groups, though students in the lower two groups grew slightly more than students in highest group from pretest to final test.

The data analyzed for this example come from the evaluation study. The subsample considered here consists of item responses from 293 4th and 5th grade students who received the new curriculum from four assessments: a pretest in mid-September, an interim test in mid-October, a posttest in early December, and a final end-of-year test in mid-May. The assessment data being analyzed consists of responses to the 18 common items (the same on all four assessments). The assessments were designed such that about half directly correspond to tasks from the LMR curriculum and half are drawn from other sources, primarily the regular math textbook (which is the same in all classrooms). Dichotomous scoring was used for all items.

Here, we demonstrate how we can use the proposed models for differential growth to examine the relationship between growth and prior achievement in more detail than was done by Saxe et al. (2013). By using the proposed latent variable model, we avoid any bias caused by measurement error in the observed scores and are able to consider simultaneously both components of initial status discussed above. We focus on the students in the LMR group because for that group we can clearly define (1) an initial time in relation to new curriculum and (2) research questions relating to entry knowledge and growth under the new curriculum.

### 4.4.2 Analysis

The analysis of the LMR data had two purposes: (1) to compare the results from models that reflect different conceptions of initial status and (2) to illustrate the interpretation of results for differential growth from these models.

Four models for differential growth were fit to the LMR data:

1. The model with slope regressed on intercept (Equation 4.17, Figure 4.3b)
2. The model with slope regressed on achievement at initial time (Equation 4.19, Figure 4.3d)
3. The model with slope regressed on residual at initial time (Equation 4.18, Figure 4.3c)
4. The general model with slope regressed on intercept and residual at initial time (Equation 4.20, Figure 4.4)

Estimation of all models was done (1) using the gllamm command (Rabe-Hesketh et al., 2004b) in Stata12 (StataCorp, 2011) and (2) using Mplus (Muthén & Muthén, 2010). gllamm used adaptive Gauss-Hermite quadrature (Rabe-Hesketh et al., 2005) with eight integration points while Mplus used adaptive rectangular (i.e. Cartesian) quadrature with 10 integration points.
4.4.3 Results

Table 4.2 contains the results from fitting each of the four models to the LMR data using both Stata and Mplus.
Table 4.2: Estimated growth coefficients and variances (with SEs) from fitting four models to the LMR data, with the slope regressed on (1) the Intercept, (2) the Initial achievement, (3) the Residual at the initial time, and (4) both the intercept and residual at the initial time (General).

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Initial</th>
<th>Residual</th>
<th>General</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gllamm</td>
<td>Mplus</td>
<td>gllamm</td>
<td>Mplus</td>
</tr>
<tr>
<td><strong>Growth Coefficients</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope ($\gamma_{10}$)</td>
<td>0.744</td>
<td>0.752</td>
<td>0.750</td>
<td>0.755</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.028)</td>
<td>(0.028)</td>
<td>(0.028)</td>
</tr>
<tr>
<td>Slope x Inter ($\gamma_{11}$)</td>
<td>0.147</td>
<td>0.139</td>
<td>0.160</td>
<td>0.149</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.035)</td>
<td>(0.034)</td>
<td>(0.034)</td>
</tr>
<tr>
<td>Slope x Resid ($\gamma_{12}$)</td>
<td>0.160</td>
<td>0.149</td>
<td>1.234</td>
<td>1.356</td>
</tr>
<tr>
<td></td>
<td>(0.034)</td>
<td>(0.034)</td>
<td>(0.194)</td>
<td>(0.184)</td>
</tr>
<tr>
<td><strong>Growth Variances</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$sd(\epsilon)$</td>
<td>0.321</td>
<td>0.315</td>
<td>0.346</td>
<td>0.311</td>
</tr>
<tr>
<td></td>
<td>(0.046)</td>
<td>(0.039)</td>
<td>(0.046)</td>
<td>(0.041)</td>
</tr>
<tr>
<td>$sd(u_0)$</td>
<td>1.022</td>
<td>1.104</td>
<td>1.018</td>
<td>1.082</td>
</tr>
<tr>
<td></td>
<td>(0.052)</td>
<td>(0.094)</td>
<td>(0.053)</td>
<td>(0.039)</td>
</tr>
<tr>
<td>$sd(u_1)$</td>
<td>0.248</td>
<td>0.230</td>
<td>0.243</td>
<td>0.241</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.031)</td>
<td>(0.037)</td>
<td>(0.029)</td>
</tr>
<tr>
<td><strong>Estimation Time (hours)</strong></td>
<td>7.2</td>
<td>0.35</td>
<td>59.6</td>
<td>0.45</td>
</tr>
</tbody>
</table>
The results are similar between **gllamm** and Mplus. The greatest difference between the two software packages was estimation time. The analysis took much longer in **gllamm** than in Mplus. For the general model, **gllamm** took 68 hours (almost three days) while Mplus took only 48 minutes. In terms of convergence, the most concerning difference between software is for the model with the slope regressed on the residual at the initial time. For this model, the estimation in **gllamm** converged without a problem but the estimation in Mplus converged to a boundary case (the residual variance of the slope went to zero). This persisted when different starting values were used to try to pull the estimate away from the boundary.

Using either software package, the general model with slope regressed separately on the intercept and on the residual at the initial time fit the data better than any of the more restricted models. This can be assessed in two ways; we consider significance at the 5% level. Using likelihood ratio tests to compare each of the more restricted models to the general model, which results in chi-squared distributions with one degree of freedom, the general model fit better than the others. Using the Wald test for the coefficients, $\gamma_{11}$ and $\gamma_{12}$ were both significantly different from 0 and from each other.

If we had only fit the model with the slope regressed on the intercept or on the achievement at the initial time, we would have found a small but statistically significant effect of initial status on rate of change. This would have been interpreted as students with a higher expected or actual initial achievement, respectively, being predicted to learn more under the **LMR** lessons. These results would appear to contradict the finding from Saxe et al. (2013) that students in the lower pretest score groups grew slightly more. However, we note that neither of these models provided the best fit to the data.

From the results of fitting the general model to the **LMR** data, which provided the best fit, we see that we need to separately consider the effects of the intercept and the residual on student growth. The rate of growth is predicted more by the residual at the initial time rather than the intercept. There is smaller residual variance for the slope in the models where it is regressed on the residual, indicating that the residual explains more of the variation in growth than the intercept.

To compare the magnitude of the effects of the intercept and the residual on growth, we should not use the estimates directly. Instead, because these effects enter the model as loadings for the latent variable, we need to consider the estimated regression coefficient of each in conjunction with its estimated variance. We multiply the estimated coefficient for the latent variable by its estimated standard deviation. For each one standard deviation increase, the intercept predicts a 0.09 logit increase in the slope while the residual predicts a 0.35 logit increase. Figure 4.8 illustrates this finding for the intercept (left panel) and the residual (right panel). Each plot shows the expected growth for students who are one (dashed lines) or two (dotted lines) standard deviations above or below the mean at pretest (solid line). For the intercept (Figure 4.8a), there is large variation in expected achievement at pretest but a small effect, leading to lines that appear almost parallel and therefore increase the expected variation at final test only slightly. In contrast, for the residual (Figure 4.8b), there is smaller variation at pretest around a given expected achievement but a larger effect, leading to lines that fan out and result in expected variation at final test that is almost the
same as the variation due to differing intercepts.

To make the interpretation even more concrete, we can consider the expected growth trajectories of specific students. Figure 4.9 shows the predicted growth trajectories and achievement at each time point for six illustrative students. These students were selected in three pairs with expected achievement at pretest (i.e. intercept) of approximately -1, 0, and 1; within each pair, one student (in gray) had a large positive estimated residual at the initial time while one student (in black) had a large negative estimated residual. Within each pair, we see that the expected trajectories of the students with negative residuals (dashed black lines) are flatter than the expected trajectories of the students with positive residuals (solid gray lines). As a result, the students with positive residuals have higher predicted achievement at final test than the students with negative residuals.

4.5 Discussion

This paper proposed a longitudinal item response model for differential growth based on initial status. The model was designed to answer research questions regarding the relationship between a student’s initial proficiency and his or her growth in proficiency over time. Such research questions examine for whom an instructional sequence or educational program is effective and whether the instruction or program is expected to narrow or widen an existing achievement gap.

The proposed model encompasses different conceptions of initial status that can be examined simultaneously. The model includes a more general specification of initial status,
the expected achievement at the initial time, in addition to a more specific alternative, the
residual at the initial time. The proposed model thus allows a more nuanced examination of
the relationship between initial status and rate of change than previous models that defined
initial status using only the expected achievement. If the two are found to be similar, initial
status can be interpreted straightforwardly as achievement at the initial time. If the two are
different, considering the two separately gives a more complete picture of the relationship
between initial status and growth. We can uncover whether growth is predicted by factors
common across the assessments or by factors specific to the assessment at the initial time.

A well-defined initial time with variation in status at that time is required for the use of
the proposed model to be sensible. If there is no variation in status at the selected initial time,
no relationship between status and change exists. Statistically, this is a boundary condition
for which the model is not identified and convergence issues arise for the estimation. If
there is no reason to prioritize a specific occasion as the initial time for all individuals, it
do not make sense conceptually to discuss a single relationship between status and growth,
because the relationship will change depending on which occasion is labeled as time zero.
This indeterminacy opens up an interesting possibility for an extension to the model: to
incorporate the effect of current status at multiple occasions on any subsequent growth.

There are additional important extensions to the model. The growth model presented
here was purposefully simple in order to focus on the different conceptions of initial status.
The model could also incorporate complications that are reflected in previous models for
longitudinal growth (see, e.g., Bryk & Raudenbush, 1987; McArdle & Epstein, 1987; Rau-
denbush, 1989; Willett & Sayer, 1994; Muthén & Curran, 1997). For example, the model
could be modified to include a more realistic model for change over time than linear growth.
The loadings for the slope latent variable could be estimated to obtain an empirical curve. Alternatively, additional fixed or random effects (i.e. latent variables in SEM terminology) could be added to model quadratic or higher-order polynomial growth. As a second example, the model would be extended to include additional covariates to explain variation in growth. These covariates could enter the model in a number of ways. They could directly predict the random intercept, slope(s), or the time-specific achievement. Alternatively, they could directly predict the item responses, resulting in differential item functioning, or the residuals, resulting in heteroskedastic variance at different levels. A third example would be including the proposed growth model as part of a larger model that jointly models the growth process and its effect on another process, such as an additional growth process or a survival process (cf. Curran, Harford & Muthén, 1996; Wulfsohn & Tsiatis, 1997).

One extension of particular importance is when the covariate being added represents assignment to different treatment and comparison groups. The model could then be used to examine treatment effects defined in a number of different ways, including standard effects such as different final status or different overall rates of growth and also more nuanced effects regarding different relationships between initial status and growth (Muthén & Curran, 1997; Khoo, 2001; Choi & Seltzer, 2010). The model with this extension would address the consequences of particular curricula or interventions exacerbating or alleviating pre-treatment differences in status.
4.A Appendix

4.A.1 gllamm syntax

The general model for the empirical application is written using gllamm syntax as:

```plaintext
/*
Data in long format contains the following variables:
  person   --  person id number
  occ      --  occasion number
  time     --  occasion number recoded so initial time is 0
  y        --  item responses, dichotomous
  it1-it18 --  indicator variables for the items
*/

/* Set the number of items */
global I = 18

/* Create indicator variables for initial time and not initial time */
gen t0 = time == 0
gen t_gt0 = time > 0

/* Create a variable of all 1s (for the constant) */
gen one = 1

/* Define the linear predictors for the latent variables */
eq occt: t_gt0
eq occ0: t0 time
eq int: one time
eq slop: time

/* Define constraints */
constr def 1 [occ1_1]t_gt0 = [per2_1]t0 /* variance of epsilon */
constr def 2 [per3_2_1]_cons = 0 /* uncorr slope and inter */

/* Run the model in gllamm */
gllamm y it1-it$I time, nocons link(logit) fam(binom) ///
i(occ person person) nrf(1 1 2) nip(8) adapt ///
eqs(occt occ0 int slop) constr(1 2) trace
```
4.A.2 Mplus syntax

The general model for the empirical application is written using Mplus syntax as:

```
TITLE: Model with slope regressed on intercept and error at initial time
DATA: FILE IS LMR_LatentGrowth.dat;
VARIABLE: NAMES ARE pre1 pre2 pre4 pre6 pre8a pre8b pre9 pre11 pre12
         pre14 pre16 pre17 pre19 pre20 pre22 pre24 pre26 pre27
         inter1 inter2 inter4 inter6 inter8a inter8b inter9 inter11 inter12
         inter14 inter16 inter17 inter19 inter20 inter22 inter24 inter26
         inter27
         post1 post2 post4 post6 post8a post8b post9 post11 post12
         post14 post16 post17 post19 post20 post22 post24 post26 post27
         final1 final2 final4 final6 final8a final8b final9 final11 final12
         final14 final16 final17 final19 final20 final22 final24 final26
         final27;
CATEGORICAL ARE pre1 pre2 pre4 pre6 pre8a pre8b pre9 pre11 pre12
         pre14 pre16 pre17 pre19 pre20 pre22 pre24 pre26 pre27
         inter1 inter2 inter4 inter6 inter8a inter8b inter9 inter11 inter12
         inter14 inter16 inter17 inter19 inter20 inter22 inter24 inter26
         inter27
         post1 post2 post4 post6 post8a post8b post9 post11 post12
         post14 post16 post17 post19 post20 post22 post24 post26 post27
         final1 final2 final4 final6 final8a final8b final9 final11 final12
         final14 final16 final17 final19 final20 final22 final24 final26
         final27;
MISSING ARE .;

ANALYSIS: ESTIMATOR = ML;
INTEGRATION = 12;

MODEL:
!define latent variable at each time, fix loadings to 1
f0 BY pre1-pre27@1;
f1 BY inter1-inter27@1;
f2 BY post1-post27@1;
f3 BY final1-final27@1;

!constrain variance to be equal
f1 f2 f3 (1);

!set variance at initial time to 0
```
f000;

!define latent variable for residual at time 0, fix loading to 1
f00 BY f0;
!set correlation of residual and slope to 0
f00 WITH i@0;

!contrain threshold across time
[pre1$1 inter1$1 post1$1 final1$1] (2);
[pre2$1 inter2$1 post2$1 final2$1] (3);
[pre4$1 inter4$1 post4$1 final4$1] (4);
[pre6$1 inter6$1 post6$1 final6$1] (5);
[pre8a$1 inter8a$1 post8b$1 final8a$1] (6);
[pre8b$1 inter8b$1 post8a$1 final8b$1] (7);
[pre9$1 inter9$1 post9$1 final9$1] (8);
[pre11$1 inter11$1 post11$1 final11$1] (9);
[pre12$1 inter12$1 post12$1 final12$1] (10);
[pre14$1 inter14$1 post14$1 final14$1] (11);
[pre16$1 inter16$1 post16$1 final16$1] (12);
[pre17$1 inter17$1 post17$1 final17$1] (13);
[pre19$1 inter19$1 post19$1 final19$1] (14);
[pre20$1 inter20$1 post20$1 final20$1] (15);
[pre22$1 inter22$1 post22$1 final22$1] (16);
[pre24$1 inter24$1 post24$1 final24$1] (17);
[pre26$1 inter26$1 post26$1 final26$1] (18);
[pre27$1 inter27$1 post27$1 final27$1] (19);

!define linear growth model
i s | f000 f101 f202 f303;

!regress slope on intercept and residual
s ON i f00;

OUTPUT: SVALUES TECH1;
References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


StataCorp (2011). *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.


