Evaluating the Equal-Interval Hypothesis with Test Score Scales

Benjamin W. Domingue

University of Colorado at Boulder, benjamin.domingue@colorado.edu

Follow this and additional works at: https://scholar.colorado.edu/educ_gradetds

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

Domingue, Benjamin W., "Evaluating the Equal-Interval Hypothesis with Test Score Scales" (2012). School of Education Graduate Theses & Dissertations. 19.

https://scholar.colorado.edu/educ_gradetds/19

This Dissertation is brought to you for free and open access by School of Education at CU Scholar. It has been accepted for inclusion in School of Education Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Evaluating the Equal-Interval Hypothesis with Test Score Scales

by

Benjamin Webre Domingue

B.S., University of Texas at Austin, 2001
M.A., University of Texas at Austin, 2006

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
School of Education

Department of Research and Evaluation Methodology

2012
This thesis entitled:
Evaluating the Equal-Interval Hypothesis with Test Score Scales
written by Benjamin Webre Domingue
has been approved for the School of Education
Department of Research and Evaluation Methodology

Derek Briggs

Dr. Greg Camilli

Dr. Finbarr Sloane

Dr. Jason Boardman

Dr. Robert McNown

Date __________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
In psychometrics, it is difficult to verify that measurement instruments can be used to produce numeric values with the desirable property that differences between units are equal-interval because the attributes being measured are latent. The theory of additive conjoint measurement (e.g., Krantz, Luce, Suppes, & Tversky, 1971, ACM) guarantees that interval scales are possible—for latent and manifest variables alike—if certain axioms hold. However, ACM was initially developed under the assumption that data could be gathered to test the axioms that was free from measurement error. It wasn’t until Karabatsos (2001) that the methodology allowed for measurement error. In this dissertation, an improved version of Karabatsos’s methodology is applied to simulated and empirical data to test whether such data are consistent with the axioms.

It is first shown that the methodology behaves reasonably using data simulated to meet the cancellation axioms of ACM. It is then shown that the methodology is capable of distinguishing data simulated to meet the axioms from data that is not. In particular, it is demonstrated that the methodology is sensitive to item-side violations of the axioms. Empirical examples are then used to illustrate the fact that test score data may or may not conform to the ACM axioms. Empirical demonstration shows that an existing test scale thought to satisfy the ACM axioms using the Karabatsos (2001) approach does not do so using the modified approach here.

Since not all data may meet the ACM axioms (and hence not warrant interval interpretations), this dissertation also examines whether scale distortions can lead to erroneous conclusions. At the score-level, an approach was developed using “difference matrices” to highlight the fact that when the Rasch (1960) model is applied to certain non-Rasch data, the estimates will be more compressed at lower abilities. This same phenomena was noted in two simulations meant to capture how educational assessment data is used—with respect to schools and educational interventions—
although the effects of the distortions were small.
Dedication

Dedicated to my family. And brevity.
Acknowledgements

I’d like to acknowledge my academic mentors from the last several years. Conveniently, they all happen be on my committee. For the combination of encouragement and critical feedback, I’ll always be thankful. Especially Derek, without whom this dissertation would be both much shorter and much worse.

I’d also like to acknowledge the steadfast support of my alarm clock.
## Contents

### Chapter

1. **Introduction**
   - 1.1 Research Questions .............................. 6
   - 1.2 Overview of Chapters ............................ 7

2. **Background**
   - 2.1 In educational measurement, what does measurement mean? ................. 8
     - 2.1.1 Classical Measurement .......................... 8
     - 2.1.2 Operationalism ................................. 11
     - 2.1.3 Representational Measurement .................. 12
     - 2.1.4 Realism ........................................ 14
   - 2.2 Item Response Theory .............................. 15
     - 2.2.1 The Rasch model as a philosophically distinct tradition ............... 17
   - 2.3 Terminology ....................................... 18
   - 2.4 Additive conjoint measurement ........................ 19
   - 2.5 Are Rasch measurement and ACM compatible? .............................. 26
     - 2.5.1 The role of error ................................. 26
     - 2.5.2 Philosophical distinctions between RM and ACM ...................... 29

3. **Methodology** ........................................ 32
   - 3.1 Literature Review .................................. 33
3.2 Karabatsos’s Methodology ........................................... 37
   3.2.1 The Metropolis-Hastings algorithm .............................. 40
   3.2.2 Single Cancellation ............................................. 41
   3.2.3 Double Cancellation ............................................ 43
   3.2.4 A brief semantic digression .................................... 44
3.3 Shortcomings of the original approach ............................. 46
   3.3.1 Order along the minor diagonal .............................. 46
   3.3.2 Are Karabatsos’s ordering restrictions the right ones? .... 47
3.4 An omnibus approach ................................................. 49
   3.4.1 Conceptual Overview ........................................... 50
   3.4.2 Technical Specification ....................................... 52
3.5 The Fit of the Rasch Model .......................................... 55
4 Simulation .................................................................. 60
   4.1 Basic Simulation ................................................... 61
      4.1.1 “Size” Analysis ............................................... 61
      4.1.2 “Power” Analysis ............................................. 64
   4.2 Reflections on first simulations ................................. 67
   4.3 Second Simulation .................................................. 69
      4.3.1 Number of sampled 3-matrices ............................ 70
      4.3.2 Results ......................................................... 71
      4.3.3 Comparing fit statistics to the checks ..................... 77
   4.4 Difference Matrices ................................................ 84
   4.5 Second simulation Redux ......................................... 89
      4.5.1 Difference Matrices, Revisited ........................... 92
      4.5.2 Quality of sum scores as indicators of ability .......... 95
   4.6 Effect of Distortions ............................................. 96
4.6.1 Treatment Simulation .............................................. 96
4.6.2 Clustering Simulation ............................................. 100
4.7 Discussion .............................................................. 104

5 Empirical Demonstration ........................................... 110
5.1 Lexile Scale ............................................................. 110
  5.1.1 Background on Lexile framework ............................. 111
  5.1.2 Data ................................................................. 112
  5.1.3 Replication .......................................................... 113
  5.1.4 Results ............................................................ 114
5.2 State Data ............................................................... 123
  5.2.1 Data ................................................................. 126
  5.2.2 Results ............................................................ 127
5.3 Conclusions ............................................................. 133

6 Discussion ............................................................... 135
6.1 Limitations ............................................................. 142
6.2 Future Research ....................................................... 144
6.3 Implications for Practice .......................................... 144
6.4 Closing Remarks ..................................................... 145

References .............................................................. 148

Appendix

A Methodology Revisions ............................................ 154
  A.1 Double Cancellation, again .................................... 154
  A.2 Updated Methodology .......................................... 157
    A.2.1 Technical Specification ................................... 158
A.3 Results ........................................................................................................ 158
   A.3.1 A comparison of old and new methodologies .................................. 158
   A.3.2 Simulation results ............................................................................. 159
A.4 Why the updated methodology doesn’t affect the key results from the dissertation . 160
### Tables

**Table**

1.1 Comparison of Decisions based on Rasch fit statistics and the checks .............. 5  

2.1 Terminology Examples ................................................................. 20  

4.1 Summary table for second simulation: Mean % of detected distortions as a function  
of the simulation conditions. .................................................................. 74  

4.2 Item Parameters from iteration of simulation with 40 items with non-zero guessing  
parameter and 40 items with non-unity discrimination ............................... 80  

5.1 Item indicators for Duval data .................................................................. 117  

5.1 Item indicators for Duval data .................................................................. 118  

5.1 Item indicators for Duval data .................................................................. 119  

5.2 Fit statistics for state: Quantile of empirical fit statistic relative to a distribution of  
simulated fit statistics ............................................................................. 132  

5.3 Comparison of Decisions based on Rasch fit statistics and the checks .............. 134  

6.1 Comparison of Decisions based on Rasch fit statistics and the checks .............. 140  

6.2 NAEP Long-term trend results .................................................................. 146  

A.1 Truth tables for double cancellation statements ......................................... 156
Figures

Figure

2.1 Rendering of Row and Column Ordering .................................................. 24

3.1 Rendering of Row and Column Ordering .................................................. 38

4.1 Proportion of violations for items at each sum score when underlying model is truly
Rasch ................................................................. 63

4.2 Proportion of violations for items at each sum score when guessing is varying .... 66

4.3 Proportion of violations for items at each sum score when discrimination is varying . 68

4.4 Stability of the proportion of checks showing distortion over varying numbers of
draws. Vertical red line represents 15,000 draws while the x-axis extends all the way
to 100,000 draws. Y-axis shows percentage of violations detected for each item. Note
that, in general, after the lines cross the vertical red line they have flat slopes....... 72

4.5 Performance of checks in simulation iterations with same number of items that have
guessing and differential discrimination. Black lines indicate truly Rasch items while
red lines indicate items that have either guessing or non-unity discriminations...... 76

4.6 Small Multiple Figure comparing checks to fit statistics ............................. 79

4.7 Comparison of Checks to Item Q-index. Black dots represent Rasch items while red
dots are items with some deviation from the Rasch model .............................. 83

4.8 Effect on $\Delta'$ and $\Delta''$ for case of uniform stretching. $\Delta'$ is solid, $\Delta''$ is dashed .. 86
4.9 LOESS curves for $\Delta'$ when difference matrices of ability estimates (black-Rasch, red-3pl) are compared to true differences matrices. ......................... 87

4.10 LOESS curves for $\Delta''$ when difference matrices of ability estimates (black-Rasch, red-3pl) are compared to true differences matrices. ......................... 88

4.11 Scatterplots of abilities for secondary simulation in case of Rasch items .......... 91

4.12 Mean absolute deviations between abilities and true-Rasch abilities. Black represents Rasch and red represents 3pl. ........................................ 91

4.13 LOESS curves for $\Delta'$ when difference matrices of ability estimates (black-Rasch, red-3pl) are compared to true-Rasch differences matrices. ......................... 93

4.14 LOESS curves for $\Delta''$ when difference matrices of ability estimates (black-Rasch, red-3pl) are compared to true-Rasch differences matrices. ......................... 94

4.15 Regression lines from regressing post- to pre-test difference on the pre-test score interacted with a dummy variable indicating treatment. The black line are the predicted differences for the control group and the red line represents the fitted differences for the treatment group. At top, the data is purely Rasch and the constant treatment effect is roughly recovered (note that the interaction p-value is relatively large. At bottom, the items do not conform to the Rasch model and this suggests, incorrectly, that lower ability groups receive less of an effect from the intervention than higher ability groups. ........................................ 99

4.16 Scatterplots comparing the abilities on the x-axis (standardized true abilities in black, standardized Rasch abilities in red) to the standard deviation of each quantity on the y-axis. The similarly colored curves are LOESS curve fitted through each set of points. Each dot represents a distinct school mean ($N = 50$). ......................... 103

5.1 Proportions of distortions based on Lexile ordering for Duval data. .................. 115
5.2 Fit Statistic Histograms: Solid black curve gives density curve of fit statistics for Duval data. Solid vertical lines show mean and plus/minus one standard deviation from the mean for this empirical density. The dotted lines are the mean and plus/minus one standard deviation for fit statistics based on Rasch data simulated to be similar to the Duval data. The fact that nearly all of the empirical distribution is outside of this range suggests that the Rasch model is not a plausible data generating model. 120

5.3 Small multiples for Duval data ................................................................. 122

5.4 Item Parameters: Analysis of 3pl parameters estimated from supposedly Rasch Duval data. ................................................................. 124

5.5 Item Parameters as compared to the proportion of failed checks. ............... 125

5.6 Results of Checks for state data ................................................................. 128

5.7 Small Multiple for state showing checks versus fit statistics ......................... 131

A.1 Comparison of 95% credible region for posterior from updated and original methodology. Colored circles represent mean of posterior under either method. Larger circles represent observed proportions of correct responses for the 3-matrix (note the violation for cell (3, 2)). Dotted vertical lines indicate “corner” cells. .................. 161

A.2 Demonstration of ability of updated methodology to detect items that deviate from Rasch model via either a non-unity discrimination or non-zero guessing. .................. 162
Chapter 1

Introduction

Man is not a rational animal, he is a rationalizing animal.

—Robert A. Heinlein, Assignment in Eternity (1953)

In an attempt to rationalize educational achievement as measured by scale scores, it seems natural to analogize educational achievement with height. Measuring length is a measurement process we're all accustomed to and is certainly far more intuitive than “measuring” educational achievement. One particularly convenient feature of how we measure length is that the scale is interval: growth of an inch for a person that is 5 feet tall means the same amount of change in height as growth of an inch for a person that is 6 feet tall. The simplicity of measuring length (and the fact that this measure has a number of useful properties) should not be taken as indication that all measurement is so straight-forward. In particular, the measurement of psychological variables that might be associated with educational achievement (more specifically, reading or math ability) is a subject fraught with philosophical and technical difficulties. This dissertation grapples with several questions related to the measurement properties of educational scales: do data from large-scale assessments generate scales that are interval? how do we know? what happens when scales are not interval?

If interval interpretations of educational scales are possible, this would make it far easier to communicate information about student's growth to parents. People have an intuitive grasp of what it means to grow in the context of height. Furthermore, parents typically understand rather complex questions about growth, such as whether a certain amount of absolute growth
(say 3 inches) is “a lot” or “a little” given a certain context (amongst 5 year old males perhaps). This understanding could be called upon when educators attempt to communicate information about student’s ability, and growth in that ability. For example, if a student who has grown equal amounts in her mathematical ability over the last 3 years suddenly grows only half that amount in the present year, a parent would be naturally inclined to ask questions about whether this deceleration is normal or worrisome.

Being able to place student achievement on an interval scale would also allow for a wider variety of quantitative analyses. Under the theory of Stevens (1946), a broader range of analyses can be conducted with interval data as compared to ordinal data. The simple act of taking a mean is permissible with interval data but is not considered a meaningful operation to perform upon purely ordinal data. A more complex analysis might involve the use of scale scores in some proposed educational accountability systems. Many value-added models, for example, use differences between a student’s actual score and their expected score. If score scales are interval then this is a reasonable procedure. If not, then this approach is questionable since differences across the score scale are not consistently interpretable as equal changes in underlying ability.

Since there is so much to gain from interval scales, it seems that some have simply assumed that educational scales are interval.

This vertical scale can be viewed as a developmental continuum. As students develop new capabilities, they move along the continuum, as demonstrated by their scale score. Scale scores are units of a single, equal-interval scale applied across all levels of TerraNova regardless of grade or time of testing. These scores are expressed in numbers that range theoretically from 0 to 999. The equal-interval property of scale scores permits arithmetic functions to be performed using the scale scores. CTB—McGraw Hill, 2001, p. 322

There is little, besides wishful thinking, underlying such a claim. There is nothing inherent to the methodology, item response theory, used to create the scale that ensures the equal-interval properties described above. Consider the following quote from a foundational work in item response theory.

At various times in this book, however, we shall treat a measurement as having
interval scale properties and the theory underlying it yield only a nominal or, at best, an ordinal scale....from a pragmatic point of view, the only meaningful evaluation of this procedure is one based on an evaluation of the usefulness of the resulting scale. If we construct a test score by counting up the correct responses and treating the resulting scale scores as having interval properties, the procedure may or may not produce a good predictor of some criterion. To the extent that this scaling produces a good empirical predictor the stipulated interval scaling is justified. Lord & Novick, 1968, p. 20-21

This is guidance for when a scale may be treated as interval, not a rule about when a scale actually is interval. This rule of thumb depends solely upon utility and correlation with some external criteria. Nothing here depends on the quality of the assessment used to generate the data being scaled. There are easier and harder measurement challenges in educational assessment (for example, accuracy of spelling in the third grade versus the measurement of the full range of mathematical ability over grades 3–9). Presumably this leads to some natural variation in the quality of the resulting assessments. There is probably similar variability in the size of the intellectual “leap” involved in assuming that various scales are interval. The requirements for an equal-interval scale may be far more challenging to meet with some data than others.

There are a number of potential problems associated with assuming an interval scale when one does not exist. Going back to the context of value-added models, the differencing that occurs—between student’s actual and expected scale scores—would appear to require an interval scale. One could claim, consistent with Lord and Novick (1968), that value-added models are useful, in the sense that they allow one to do something and hence the operation is warranted. Ballou (2009) notes that such an assumption could be problematic. Imagine a scale where, at the low end, scores are quite compressed (a small score difference represents a larger difference in ability when compared to the top end). If a teacher was to be evaluated by a value-added model in such a school, it may well matter where their students were on the scale since small movements by students on the high and low ends of the scale might imply different things about changes in student achievement. A teacher unfairly penalized by such a model would presumably have a dim view of the “utility” of
the equal-interval assumption. A study of the potential effects of wrongly assuming interval scale scores is undertaken in Chapter 4.

It is difficult to verify that a scale for measuring psychological variables, such as math or reading ability, is interval. But is is not impossible. The development of additive conjoint measurement by Luce and Tukey (1964) offered a framework that could be used to evaluate whether a manifest or latent variable has a quantitative structure that would support equal-interval scaling. Conjoint measurement deals with quantities that depend upon at least two factors. Density depending upon weight and volume is one example. Another example is a response to a test item which, in the Rasch (1960) formulation, depends upon student ability and item difficulty. Conjoint measurement proves that if certain sets of axioms are true, then a set of scales (in the context of test scores: scales for student ability, item difficulty, and the quantity that results from their combination) are guaranteed to exist although the theory does not provide a proscriptive method for their construction. An important point is that the measurement axioms require certain orderings to hold with respect to the two quantities under examination. This is a problem since additive conjoint measurement relies upon comparisons between fixed, static quantities and item response theory is expressed entirely in terms of variables assumed to be measured with error. This was partially resolved in Karabatsos (2001), although as I detail in Chapter 3 there were some flaws in that approach that I correct here.

In this dissertation, I build upon the ideas of Karabatsos (2001) and Kyngdon (2011) in extending a method for evaluating whether a collection of dichotomously scored test item responses can bear the weight of an interval scale. My methodological extensions are two-fold. First, I make the corrections just noted to the approach. Second, I demonstrate that initial applications of this method may have been insufficient for the stated objective of testing the axioms. Using the modified methodology, I will be able to analyze whether empirical data seems to allow for an assumption of

---

1 Some might argue that this is being somewhat unfair to Lord and Novick (1968) in that I am not taking into account their criteria of predictive validity. However, this doesn’t seem all that useful a guide. If value-added estimates computed on the basis of a non-interval scale correlated at 0.3 with a measure based on classroom observation, would that be sufficient? Would accurate causal inferences result? Evaluating teacher effectiveness is a topic fraught with difficulties that I shall not attempt to go into, but it is hard to see how mistakenly claiming an ordinal scale is interval is going to improve anything.
underlying interval quantities. The “checks” (as I call them) analyze whether the axiom-required orderings, of both individuals and items, hold across the data. To the extent that they do, this will be evidence that an interval scale is possible with such data. In Chapter 4, I also perform extensive tests of the checks via simulation that demonstrate they are reasonable statistical tools.

As noted earlier, the theory of conjoint measurement does not produce a candidate interval scale. This is another matter, although obviously quite an important one. One plausible method for creating such a scale is the Rasch (1960) model. It is important to note that Rasch measurement is somewhat distinct from the broader field of item response theory. From the perspective of conjoint measurement in particular, this model has the important feature of meeting the axioms of conjoint measurement when it is true (Brogden, 1977, Borsboom, 2005, Borsboom & Scholten, 2008, Michell, 2008b), something not necessarily true for more elaborate models. As a consequence, there will be a lot of attention to the fit of the Rasch model when it is used for scale creation. Much research has been done over the years on fit statistics for the Rasch model. It is important to note that these statistics address a separate, indeed subsequent, question from the one that is at the heart of this dissertation. Assuming that there is some latent variable underlying test response data, I begin by asking whether data satisfy the checks. If so, then conjoint measurement implies interval scales are possible. After the analysis using the checks, it may then make sense to ask whether the Rasch model creates leads to an interval scale. Rasch fit statistics provide important information about whether the stricter assumptions of the Rasch model are reasonable for a given set of data.

<table>
<thead>
<tr>
<th>Rasch Fit Statistics</th>
<th>RM fits</th>
<th>RM does not fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Checks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>consistent with ACM</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>not consistent with ACM</td>
<td>IV</td>
<td>III</td>
</tr>
</tbody>
</table>

Table 1.1: Comparison of Decisions based on Rasch fit statistics and the checks
To be very clear about this contrast, consider Table 1.1. Cell I would be a best case scenario. Data in this cell is consistent with both the checks and the Rasch model (as indicated by fit statistics). The resulting Rasch scale would have interval properties. Cell IV, on the other hand, should not actually be possible if the checks are working as desired. Data generated by the Rasch model will be interval, so it should not be possible for the Rasch model to hold but for the checks to indicate violations. An insight I present in Chapter 5 is that there is an important distinction between cells II and III. Cell II contains data that supports an interval interpretation but is not well fit by the Rasch model. Although it is not a question I attempt to answer, it is natural to wonder what model may fit such data and produce the desired interval scale. Data in cell III does not seem to allow for an interval scale. Such data may be best treated as purely ordinal, although even that is an assumption that could be tested rather than simply assumed.

1.1 Research Questions

Summarizing the above discussion, I pose three specific research questions answered in this dissertation. First, are the stochastic checks suggested in Karabatsos (2001) a robust statistical tool [Q1]? As a subquestion, what, if anything, do the checks add to our knowledge of misfit when they are combined with traditional Rasch fit statistics [Q1A]. Second, do certain deviations from the assumptions of IRT models result in meaningful scale distortions [Q2]? In particular, if the Rasch model is fit to data that were generated by items whose behavior is actually more complex, how does this distort the scale? Would the scale distortions lead to mistakes when used in real world settings? Third, given that certain test score scales are based on the Rasch (1960) model, this suggests that their creators believe the data to meet the assumptions of the Rasch model. As I've noted, this implies that the scales should be interval. The question I ask is how consistent some example datasets, scaled with the Rasch model, are with the conjoint measurement axioms [Q3].

---

2 I concentrate on a relatively simple Rasch model. It is possible that more sophisticated models, such as those that allow for multidimensionality, may fit. Such models, however, would require a far more sophisticated version of the checks.

3 One alternative approach would be to assign scores in a manner known to be consistent with the axioms of conjoint measurement, such as numerical conjoint measurement (e.g., Kruskal, 1965). Studies (Nickerson & McClelland, 1984) suggest that this approach is not foolproof either.
1.2 Overview of Chapters

This dissertation proceeds as follows. In Chapter 2, I discuss key background concepts including what it means to measure along with background on item response theory and axiomatic conjoint measurement. Chapter 3 covers the analytical approach I use in this paper, an approach I will argue is an improvement of Karabatsos (2001). Chapter 4 describes the statistical properties of this approach via a number of simulation exercises. In particular I demonstrate that the approach has statistical properties that ensure it can be used reliably. I also discuss the consequences of wrongly assuming interval structure in simulations of real world uses of educational assessment data. Chapter 5 applies the updated methodology to two sets of empirical data. I demonstrate that there is a fundamental distinction between the two sets of data in terms of their placement in Table 1.1. Finally, Chapter 6 includes a review of this research as well as some limitations and notes on future research.
Chapter 2

Background

2.1 In educational measurement, what does measurement mean?

Measuring psychological variables, one instantiation of which is educational measurement, is an arduous task. To speak about it coherently, one must have in mind a definition for “measurement”. For several millennia, up until the 20th century, there was rough agreement about what this term meant. However, confusion about whether measurement was possible for psychological variables led to a radical shift amongst many social scientists about what it means to measure. Reliance on this later meaning has caused psychometrics to develop in such a way that Michell (2008b) refers to it as “pathological.” What one means by “measurement” is thus a decision of great importance. I will describe three paradigms—classical, operational, and representational—before describing the realist position laid out in Michell (2005) that I adopt.

2.1.1 Classical Measurement

Discussions of measurement date back to at least Ancient Greece. More detailed accounts can be found in Michell (1990) or Michell (1999); here, I concentrate on the basic features of the classical theory of measurement. In the classical theory, “measurement is the attempt to estimate the ratio between two instances of a quantitative attribute, the first being the magnitude measured, and the second being a known unit” (Michell, 2005, p. 287). Measuring length, as with a ruler, is the quintessential example of this. Some standard unit, be it the foot or the meter, is specified and the subsequent activity is to estimate the ratio of the object in question to that standard. I,
for example, am between 69 and 71 times as long as the standard inch. This estimation of ratio is
the sine qua non of measurement to the classicist. To the classicist, any procedure which does not
attempt this is not measurement whatever else its merits.

It is immediately apparent that there are useful non-numeric classifications one could make
about the attributes of an object. In my pantry, there is a pound of trail mix. By this, it is well
understood that I mean I would need only this much trail mix to perfectly balance a one pound
standard on a balance beam. The mass of this object is certainly a quantitative attribute.1 But
the trail mix has other attributes. It could be salty or sweet, for example. We would at first
consider these to be qualitative attributes. The quantitative imperative suggested by Michell
(1990) posits that one task of science is to “make measurable what cannot be measured” (Galileo,
as cited in Michell, 1990). We might then endeavor to measure the “flavor” of an object. To do
so in the classical sense is to hypothesize that the attribute is quantitative. This is the easy part
of the process. The difficult part is to then collect evidence in support of this hypothesis. From
this perspective, this dissertation could be framed as being about one technique for evaluating the
evidence of a certain quantitative hypothesis.

Returning to the above example of measuring height, note the role of number. The rational
numbers exist, in the classical framework, as ratios of comparable attributes of various objects.2
On the one hand, this view of numbers as existing entities is aligned with the philosophical stance
of realism that I discuss later. On the other hand, most 20th century accounts of number posit
them to be purely abstract entities. This Platonic view motivates the representational paradigm.
I explore both of these ramifications in the coming pages. As a consequence of the integral role
that number plays in the classical theory, operations upon the numbers, such as addition which
immediately corresponds with concatenation, are coherent and can be readily interpreted within
this theory. In comparison, more modern theories of number require extensive reasoning to make
sense of even the most basic operations. The most famous example of this is Whitehead and Russell

1 To be precise, a pound is a measure of weight not mass, but this does not affect the meaning of the example.
2 The integers, in contrast to the reals, can be motivated in a second, related way. For example, the number 2
“was thought to be that which is common to all pairs of things” (Michell, 1990, p. 10).
Length is an example of what is known as extensive measurement, the crucial feature of which is the ability to directly concatenate objects. An interval scale can be developed fairly directly in such cases by merely comparing objects to some pre-determined standard. In the case of length, I could cut a piece of wood into a 1 foot segment and measure anything I cared to measure by bounding the length of an object between \( n \) replications of the board and \( n + 1 \) replications. Clearly there is some error of measurement here that may be problematic in certain contexts, but I could minimize this through similar use of a 1” steel rod. Measurement of intensive quantities, which cannot be directly concatenated, is far more difficult. Indeed, it has caused scientists to rethink what they mean by measurement.

A plethora of scientific phenomena, such as temperature, do not lend themselves to direct concatenation. Psychological variables, such as ability in mathematics, also do not have this feature. Debate has long raged over whether measurement in the sense of ratio estimation is possible in such a case. The Ferguson committee (Ferguson et al., 1940) was persuaded by the view of Norman Campbell that extensive measurement is the key component of measurement and, consequently, psychological measurement is impossible. Such a view overlooked the work of Hölder, who formalized the concept of quantity (translated into English in Michell, 1996, 1997a) such that it could in principle be applied to non-extensive attributes. Hölder proved that any system of magnitudes obeying a certain set of rules is isomorphic to the real numbers, meaning that it now became possible for a researcher to conjecture that a given intensive trait is quantitative. The scientific task is then to collect evidence demonstrating that the magnitudes of the trait followed the laws laid out by Hölder. In any event, the decision of the Ferguson committee threw the field of psychological measurement into disarray. The quantitative imperative was so strong that it was unclear whether psychology could be considered scientific if it lacked measures.
2.1.2 Operationalism

Since not having measurement in psychology was virtually unthinkable, Stevens (1946) simply changed the definition. According to his famous definition, measurement is “the assignment of numerals to objects or events according to rules” (p. 667). Whereas earlier definitions of measurement prioritized the comparison of an object to some common standard (such as my 1” steel rod), a number of new operations would constitute measurement under this definition. For example, a group of people might be given a test and then “measured” by the number of correct responses. The operational philosophy of Bridgman (1927), a colleague of Stevens at Harvard, inspired this new definition of measure. The key tenet is that an attribute one wishes to measure is best defined by the operations used in the measurement process. Such a view reached its pinnacle in the definition of measurement in Dingle (1950): “any precisely specified operation that yields a number” (p. 11). In the context of educational assessment, this brings immediately to mind the intensely precise set of rules and regulations that surround most standardized testing.

Stevens (1946) did recognize that scales of measurement created via different sets of rules may have different properties. He classified scales into four rough types: nominal, ordinal, interval, and ratio. For each scale, different “permissible statistics” could be computed. If 1 represents green eyes and 0 represents blue eyes, and this is clearly a measurement scale if all that is required is a rule, then it would not make much sense to compute a mean of these two values. Such a scale is nominal and the mode would be the appropriate measure of central tendency. Interval scales allowed for the computation of a more desirable set of statistics such as the mean and standard deviation. This would possibly be a workable system if one had criteria for a ratio or interval scale as opposed to one that was only ordinal or even nominal. Instead, interval scales are largely determined by the user’s choice to compute certain statistics. This is a qualitatively different approach to measurement than the classical approach considered above.

Despite its shortcomings (primarily that few rules are in place to ensure that permissible statistics are computed only when appropriate), the definition and taxonomy for scales or measure-
ment put forward by Stevens (1946) have become the standard terminology regarding measurement in psychology (Michell, 1997b). Within large-scale assessment, I would assert that operationalism is the dominant paradigm (Michell, 2008b makes a similar claim) and the “rules” for assigning numbers typically come from the tradition of item response theory that follows from (Lord & Novick, 1968).\textsuperscript{3} Item response theory has a number of features that make the demands of large-scale assessment easier (e.g., item banking, computer adaptive testing, mixed format tests, etc.), but there is nothing in the theory that formally guarantees an interval scale. This tradition has come under heavy criticism from Joel Michell (1990, 1997b, 2005, 2008b). Perhaps the most pertinent criticism, for this study, is that the operationalist position has not encouraged the kinds of verification (indeed, what is there to verify so long as the rule has been followed?) that are at the heart of classical measurement and, as we shall now see, representational measurement.

2.1.3 Representational Measurement

Representational measurement is the most recently developed paradigm and is associated with the \textit{Foundations of Measurement} series (Krantz et al., 1971; Suppes, Krantz, Luce, & Tversky, 1989; Luce, Krantz, Suppes, & Tversky, 1990) which built upon the work of Suppes and Zinnes (1963) and Luce and Tukey (1964). There are two key elements of representational measurement for my purposes: how it reconceives number and its reliance on axiomatic frameworks. The building blocks of the representational theory are known as empirical relation systems. Empirical relation systems consist of a set of objects and a set of qualitative observations. For example, consider a set of steel rods. I could perhaps observe that rod \(A\) is longer than rod \(B\) and that if I concatenate rods \(A\) and \(B\) they are together longer than rod \(C\). I could even potentially notice that rod \(A\), when concatenated with itself, is exactly the same length as rod \(D\). This is where things get interesting. By the 20th century, it had become common for the reality of numbers, as expressed in the classical sense, to be denied and for numbers to be viewed as purely abstract entities. As such,

\textsuperscript{3} I also think it likely that most psychometricians are unaware of the historical roots of their adopted philosophy, but this is difficult to demonstrate.
they can not be used directly at this juncture since they are not real given their abstract nature. While a classicist would say that rod $D$ is twice the length of rod $A$, the representationalist, having declared rod $A$ to be the standard unit, would say that 2 is the number assigned to the length of rod $D$. The number 2 here can be used to represent the relationship between the lengths of rods $A$ and $D$. This representation of 2 will only be useful so long as the empirical relation system meets certain criteria. In the classical framework, the number 2 would be defined as the ratio of these rod’s lengths. Although being aware of this distinction is important since it is the impetus for representational measurement, I will not focus in great detail on the conflicting roles of number here (see Kyngdon, 2008b or the critique in Michell, 1990, p. 46–49 for accounts on how the role of number matters in psychometrics).

The two paradigms, classical and representational theories of measurement, are similar in that they both require arguments to be made about the quantitative hypothesis (one can not merely start working with the permissible statistics that imply quantity, such as the mean). In classical measurement, arguments were based on extensive concatenation. Representational measurement uses an axiomatic framework such that if the evidence implies certain axioms hold, then a set of theories implies the existence of interval scales (I discuss the nature of the theories in detail in subsequent sections). This method of argumentation in the representational theory is potentially useful in psychometric situations. It has been shown that when certain axioms hold in certain systems, then interval scales can be constructed based upon those systems. The prime example of this are the axioms associated with additive conjoint measurement. Additive conjoint measurement concerns systems where one variable is a non-interactive function of two other variables. For example, density is a function of mass and volume. When the qualitative observations made in the empirical relation system can be shown to be consistent with a certain set of axioms, then Luce and Tukey (1964) demonstrated that interval scales could be created for all 3 variables (density, mass, and volume in the example). This is crucial since it demonstrates one way of verification that would be potentially useful in the case of psychological measurement.
2.1.4 Realism

In this dissertation, I embrace the realist position expressed in Michell (2005). The key tenet of realism is “that a proposition is true if and only if things are as proposed” (p. 286). This commitment to truth entails embracing the existence of numbers. Consider the following: “It unfolds from this standpoint that reports of measurements (for example, the report that at such and such a time and place, Smith’s weight is 90 kg) are true if and only if things are literally as reported (i.e., at that time and place, Smith’s weight really is 90 kg)” (p. 286 Michell, 2005). For this statement—Smith’s weight is 90 kg—to be true, the number 90 must be something that actually exists. Such a stance is at odd with the representational view of numbers as Platonic entities. The classical position is philosophically consistent with realism, but does not offer the appropriate tools for demonstrating that psychological variables can be measured on an interval scale. However, while the representational world view is not consistent with the realist one, there is nothing stopping us from using their extensively developed tools.

Mechanically, this will work as follows. I will attempt to demonstrate that the axioms of conjoint measurement hold using the observed proportions for individuals responding correctly to dichotomously scored assessment items. If they do, then the theory of conjoint measurement indicates that interval scales would be possible for the given assessment. The approach I take (based on Karabatsos, 2001) is non-constructive, so I will not have a candidate scale at that point. A candidate scale will be generated via item response theory, I say more about this in the coming pages. The difficulty in this endeavor is due to error. The laws of conjoint measurement are deterministic. I work with proportions of correct responses from collections of individuals and these proportions are noisy estimates of (hypothesized) true underlying probabilities. This measurement error makes work with the deterministic axioms challenging. After describing item response theory and additive conjoint measurement, I demonstrate how these difficulties are resolved over the course of the coming chapters.
2.2 Item Response Theory

As I’ve noted, item response theory (IRT) is the main source of “rules” (in the operational sense) for creating scales. Consider a response from individual \( j \) to item \( i \) that has been coded as correct or incorrect. Representing these two response categories by 1 and 0 respectively, we can conceive of the response as a Bernoulli random variable \( Y_{ij} \) with probability of a correct response \( P_{ij} \). IRT posits that this probability is a function of item and respondent characteristics. If an entire set of items depends upon only a single latent trait, then the probability of a correct response can be parametrized as:

\[
P_{ij} = \Pr(Y_{ij} = 1) = c_i + (1 - c_i) \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}.
\] (2.1)

This particular parametrization is known as the 3 parameter logistic (3pl) model (Birnbaum, 1968). Only a single parameter relates to the individual: \( \theta_j \), commonly perceived as representing examinee \( j \)’s ability on the latent construct of interest. \( \theta \) is assumed to be a continuous quantity, with emphasis on the assumption of quantity. What is necessary is some kind of test of this assumption. One method of checking this assumption would be through model fit statistics. These may be poor measures (e.g., Karabatsos, 2000), however, and the approach that I take does not rely solely upon them.

The parameter \( b_i \) is the item difficulty parameter. If all other values are fixed, large values of this parameter correspond to smaller probabilities of correct responses. Larger values correspond to more difficult items. \( a_i \) is the discrimination parameter. This item controls the slope of the ogive. Larger values lead to more Guttman-like items (Guttman, 1941) while smaller values lead to items with little variability in the probability of a correct response for respondents across the ability continuum.\(^5\) The final parameter, \( c_i \), is known as the guessing parameter which functions as a lower asymptote for the probability of a correct response. On a multiple choice item with 4 possible responses, this model assumes that respondents are guessing with a probability of 0.25, which is the proportion of the ogive at the lower asymptote.

\(^4\) In this study I restrict attention to dichotomously coded responses. Such items form the basis for most large-scale assessments that form the basis of our current educational accountability system and considering polytomously coded responses leads to more complicated models and thus more complicated theories of measurement.

\(^5\) Guttman (1941) items are items for which respondents endorse the correct response if \( \theta_j > b_i \) and the incorrect response otherwise.
choices, we might expect to see values of around 0.25 for this parameter although estimates for this value do vary greatly in practice. One simplification of the model is to eliminate the guessing parameter, leading to the 2 parameter logistic (2pl) model. In this study, I focus primarily on the 1 parameter logistic model (1pl). It is formed by assuming that all items are equally discriminating. I also borrow liberally from the tradition of Rasch (1960) modeling, which uses the 1pl but which also emphasizes a different set of questions about psychological measurement.

There are two additional assumptions required before estimation of item parameters is possible. The first, unidimensionality, is implicit in Eqn 2.1. The response to an item is only dependent upon the single attribute captured in $\theta$. As a contrast, responses on a math test to word problems with extensive written stems may also depend upon the verbal scales of an examinee, not simply their math ability. Multidimensional item response models allow for the relaxation of this assumption exist but are not a focus of this dissertation (and are not used much in large-scale assessment). The second assumption is that, conditional on ability and the item parameters, item responses are independent. This allows the joint likelihood to be written as

$$L(\theta_j, a_i, b_i, c_i|Y_{ij}) = \Pi_i \Pi_j \left( P_{ij}^{Y_{ij}} (1 - P_{ij})^{(1-Y_{ij})} \right).$$ (2.2)

This second assumption is known as local independence. For example, if multiple items on a reading comprehension exam draw from the same passage, this may induce dependencies in the responses. I also assume local independence holds in this study.

In the 1pl case, some form (conditional or marginal) of maximum likelihood can be used to obtain estimates for the ability and difficulty parameters (see Baker & Kim, 2004 for an extensive discussion on estimation of IRT models). One convenient feature of this approach to scaling is that these estimates are on the same scale.\(^6\) Continuing to speak of the 1pl (or 2pl) model, if a person’s ability is equal to the difficulty of an item, then the person has a 50% probability of correctly answering the item and this probability increases asymptotically to unity as an individual’s ability increases.

---

\(^6\) One must remember that this scale may not be interval, so this convenience may mask more fundamental problems.
For many educational purposes, a vertical scale is desired such that student abilities across multiple grades can be compared on a common scale. Mechanically, this is usually done through the inclusion of the same items in the test booklets that students receive in multiple grades. Such an approach is known as a non-equivalent group common item design (Kolen & Brennan, 2004). For a variety of reasons (e.g., multidimensionality or item parameter drift), a single model of the data generating process may be inadequate to describe the responses of individuals to items at every administration, possibly leading to scale distortion. Values taken from such vertical scales are frequently used in secondary statistical models, such as value-added or growth models. These modeling approaches may be flawed due to multidimensionality (Martineau, 2006) or due to side-effects of the scale creation (Briggs & Weeks, 2009). However, even if these flaws did not affect a particular combination of value-added model and scale, the interval assumption would typically still be required for most uses of the vertical scale (Ballou, 2009).

2.2.1 The Rasch model as a philosophically distinct tradition

The Rasch model (1960, RM) is mathematically equivalent to the 1pl model but typically involves a different view of measurement. The distinction between the Rasch and larger IRT community has sometimes been characterized as as the former being a search for data that fits the model and the latter a search for a model that fits the data. Rasch’s work was popularized by Benjamin Wright (B. Wright, 1968, B. Wright & Panchapakesan, 1969, B. Wright & Stone, 1979). When compared to the broader psychometric community, the Rasch measurement area is characterized by an increased focus on fit (R. M. Smith, 1991, Glas & Verhelst, 1995), attention to the causes of item difficulty (Scheiblechner, 1972, G. H. Fischer, 1995, Gorin, 2006), and a more coherent view of measurement (e.g., some understanding of the three measurement paradigms laid out earlier).

What is of great interest here is the connection between the Rasch model and additive conjoint measurement. It has frequently been posited that the Rasch model is a probabilistic formulation of additive conjoint measurement (Keats, 1967, G. Fischer, 1968, Brogden, 1977, Karabatsos, 2000).
Kyngdon (2008b) disagrees with this assessment, however, and Michell (2008a) gives additional reasons one should be wary of conflating the two. Michell makes the point that merely fitting the Rasch model does not then guarantee that all the benefits of conjoint measurement appear, a stance very much in line with the approach I take. Even if the approaches to measurement are not equivalent, they are synonymous in enough respects for both Kyngdon (2008b) and Michell (2008a) to suggest that conjoint measurement may be a framework worth considering by those working with the Rasch model.

2.3 Terminology

Throughout the remainder of this dissertation, there will be frequent references to various matrices that have been formed from dichotomously coded assessment data. Here, I clearly outline the terminology that I will use throughout when referring to these various objects. I’ve also illustrated the distinctions between the various matrices in Table 2.1. By a response matrix, I mean a matrix containing the dichotomously coded responses of each individual to each item. A data matrix or conjoint matrix will mean a response matrix in which rows have been aggregated in some manner such that rows contain the proportion of correct responses from respondents at a common ability level (I will use the number of correct items endorsed by each person, their sum score, to identify common abilities, but other approaches are possible). Finally, a 3-matrix denotes a $3 \times 3$ matrix drawn from a conjoint matrix specifically for a check of double cancellation. The rows and columns selected from the data matrix to form the 3-matrix do not need to be adjacent. I shall also use some functional and set-theoretic notation. Briefly:

- $f : X \rightarrow Y$ denotes that $f$ is a function mapping elements of $X$ into elements of $Y$.
  
  Frequently, $Y = \mathbb{R}$, the set of real numbers.

- $x \in X$ means that $x$ is an element of the set $X$.

- $\forall$ reads literally as “for all”.

- $\exists$ reads literally as “there exists”.

2.4 Additive conjoint measurement

Additive conjoint measurement (ACM) describes how interval scales may be created when one variable is a non-interactive function of two other variables. I shall write this as $z_i = f(x_i, y_i)$ where $z_i \in Z, x_i \in X, y_i \in Y$, and $f : X \times Y = Z \rightarrow \mathbb{R}$. $f$ is further restricted by the fact that it is non-interactive.\(^7\) Interval scales for $X, Y$, and $Z$ will be possible if certain orderings hold with respect to specific values of the three variables.\(^8\) In the context of educational assessment, I use $f(x_1, y_1) \geq f(x_2, y_2)$ to mean that the probability of a correct response for an individual $x_1$ to item $y_1$ is larger than the probability of a correct response for individual $x_2$ to item $y_2$.

Consider the following conditions on the triple $(X, Y, \geq)$:

1. Independence (or Single Cancellation)\(^9\)
   
   If $f(x_1, y_1) \geq f(x_2, y_1)$ then $\forall y_2 \in Y f(x_1, y_2) \geq f(x_2, y_2)$.

   What is essential here is that the rows can be ordered without reference to the columns. Therefore we have an induced order on $X$ and we may unambiguously write $x_1 \geq x_2$. An identical condition must also hold for $Y$.

2. Double Cancellation\(^10\)
   
   If $f(x_1, y_2) \geq f(x_2, y_1)$ and $f(x_2, y_3) \geq f(x_3, y_2)$ then $f(x_1, y_3) \geq f(x_3, y_1)$.

   Although this axiom is seemingly abstract, it is actually a natural consequence of additivity as I show below.

---

\(^7\) This means that $f$ can be decomposed as $g(x) + h(y)$. Transformations of $f$ are permissible. For example, if $f(x, y) = xy$ then $\log xy = \log x + \log y$ yields an acceptable decomposition. Any term involving $xy$ that can’t be split means that $f$ is interactive.

\(^8\) From the representational measurement perspective, the real numbers can not yet be introduced. What will be of interest is whether certain qualitative orderings, typically expressed using $\geq$, hold. For example (following Krantz et al., 1971, p. 17), perhaps $X$ and $Y$ are temperature and humidity respectively and the combination $(x_1, y_1)$ is judged as more uncomfortable than $(x_2, y_2)$. We could then say that $(x_1, y_1) \geq (x_2, y_2)$ without having ever referenced a numerical scale. Once the axiomatic orderings are shown to hold, then real numbers can be used to represent relationships amongst the variables.

\(^9\) The independent ordering of rows and columns can be either assumed as in Krantz et al. (1971) or established as a byproduct of solvability and double cancellation as in Michell (1990).

\(^10\) Some authors prefer to distinguish between double cancellation in its deterministic, or axiomatic, form and the version that may be shown to hold in stochastic data. They refer to the stochastic version of this as bi-cancellation or weak stochastic transitivity. I will follow Karabatsos (2001) in only referring to “double cancellation” but note that there is a distinction.
### Table 2.1: Terminology Examples

<table>
<thead>
<tr>
<th>Persons</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Response Matrix**

<table>
<thead>
<tr>
<th>Persons</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,5,6</td>
<td>0</td>
<td>0.33</td>
<td>0.33</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1,4</td>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Data/Conjoint Matrix**

<table>
<thead>
<tr>
<th>Persons</th>
<th>Item 4</th>
<th>Item 1</th>
<th>Item 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,5,6</td>
<td>0.33</td>
<td>0.33</td>
<td>0.66</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>1,4</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**3-Matrix**

- These are the raw responses for persons 1–6 to items 1–5. This is only meant as a small example; the dimensions encountered in practice would be much larger.
- Note that students with common sum scores have been aggregated. In addition, the rows and columns have been ordered so that the most difficult item is on the left and the sum scores are increasing from top to bottom. The cells now also represent the group-level probability of a correct response to an item.
- Double cancellation checks occur in $3 \times 3$ matrices. For the conjoint matrix above, there would be 4 such matrices that need to be checked. The one shown results from eliminating item 3.
(3) Solvability

\[ \forall x_1 \in X, y_1, y_2 \in Y, \exists x_2 \in X \text{ such that } f(x_1, y_1) = f(x_2, y_2). \]

This axiom is necessary since ACM frequently requires the existence of an element \( x_2 \in X \) in the above for certain necessary manipulations to be possible. In the context of assessment, this theorem can be reformulated as follows: For every pair of items at different difficulties and any ability, there exists a second ability such that the probability of a correct response from a person at the first ability to the first item is equivalent to the probability of a correct response from a person at the second ability to the second item. An equivalent condition must also hold restated in terms of the existence of a second item given a pair of abilities and the difficulty of a single item.

(4) Archimedian Condition

Michell (1990) states this axiom as: “no value of a quantitative variable is infinitely larger than any other value” (p. 73). In essence, this condition ensures comparability between any two values as they can only be a finite distance apart.

The last two axioms are impossible to verify directly in most cases. Paraphrasing Michell (1990, p. 79), finding the \( x_2 \) to satisfy solvability might be impractical due to time or money constraints. Similarly, the steps required to show that a certain value is bounded (and thus restricted by the Archimedian condition) may be practically impossible. Imagine trying to demonstrate that the sun is a finite distance from the earth using a common 1 foot ruler! Scott (1964) developed some testable indirect conditions that are necessary but not sufficient conditions for 3 and 4 to hold. This study will not focus on these higher order checks since it is difficult to integrate them into the methodological framework suggested by Karabatsos (2001) and even checking the necessary (though not sufficient) conditions 1 and 2 will be a substantial step forward.

If the axioms are true\(^\text{11}\), then \((X, Y, \geq)\) is called an additive conjoint structure. It is quite surprising, perhaps, that interval scaling can be obtained as a byproduct of the above axioms but

\(^{11}\) Such definitive knowledge will not be possible in practice, of course, with assessment data.
This was shown to be the case by Luce and Tukey (1964, p. 10, see theorems VID through VIJ). It was first shown that for additive conjoint structures there exist functions $\phi_1 : X \to \mathbb{R}$ and $\phi_2 : Y \to \mathbb{R}$ (unique up to linear transformation) such that

$$f(x_1, y_1) \geq f(x_2, y_2) \Leftrightarrow \phi_1(x_1) + \phi_2(y_1) \geq \phi_1(x_2) + \phi_2(y_2). \quad (2.3)$$

This is convenient since one can now work with the simpler $\phi$ functions rather than $f$. The next step is to create dual standard sequences. Standard sequences play an analogous role in generic measurement as the role that a ruler plays in measuring length. The demarcations of inches on a standard ruler is precisely a standard sequence. Dual standard sequences are the extension of this concept to the conjoint measurement case and rely upon certain sets of points being shown to be equivalent, in terms of $f$. For $m, n \in \mathbb{Z}$, $(x_m, y_n)$ is a dual standard sequence if $f(x_i, y_j) = f(x_p, y_q)$ whenever $i + j = p + q$. A simple case will help to illustrate. If $i = q = 0$ and $j = p = 1$, then we have that $f(x_0, y_1) = f(x_1, y_0)$. This means that, in terms of $f$, there is no difference between being at $(x_0, y_1)$ and $(x_1, y_0)$. Expressed in an assessment context, a correct answer from person 0 to item 1 is just as likely as a correct answer from person 1 to item 0. Luce and Tukey (1964) then showed that if $(x_i, y_j)$ is a dual standard sequence then

$$\phi_1(x_n) - \phi_1(x_0) = n [\phi_1(x_1) - \phi_1(x_0)]$$

$$\phi_2(y_n) - \phi_2(y_0) = n [\phi_2(y_1) - \phi_2(y_0)].$$

One can now clearly see that the measures implied by $\phi_1$ and $\phi_2$ are interval. A move from $x_0$ to $x_n$ is equivalent to $n$ moves from $x_0$ to $x_1$. The distance from $x_0$ to $x_1$ is now acting as a standard unit against which other moves can be compared. Construction of the dual standard sequence may be quite challenging, but if the axioms hold then such a construction is known to exist. In psychometrics, construction of such a sequence would be difficult since it requires extensive manipulation of $X$ and $Y$ (a topic I discuss at the end of this section). In this dissertation the goal is to examine whether the data is consistent with the axioms, not trying to explicitly construct dual standard sequences.
Returning to the axioms, I want to elaborate on the double cancellation condition since it will be a focal point of this study. While it at first seems rather odd, double cancellation can be understood in the following sense. Assume that $(X, Y, \geq)$ is an additive conjoint structure. Suppose that the antecedent in the statement of double cancellation is true. This implies that

$$\phi_1(x_1) + \phi_2(y_2) \geq \phi_1(x_2) + \phi_2(y_1)$$

$$\phi_1(x_2) + \phi_2(y_3) \geq \phi_1(x_3) + \phi_2(y_2).$$

We can now combine the above inequalities to produce:

$$\phi_1(x_1) + \phi_2(y_2) + \phi_1(x_2) + \phi_2(y_3) \geq \phi_1(x_2) + \phi_2(y_1) + \phi_1(x_3) + \phi_2(y_2)$$

Cancelling terms leads to:

$$\phi_1(x_1) + \phi_2(y_3) \geq \phi_1(x_3) + \phi_2(y_1)$$

implying that $(x_1, y_3) \geq (x_3, y_1)$. So we see that double cancellation is a natural consequence of the additivity of the structure.

There is a little more intuition that one can develop about double cancellation. To show this, I refer to Figure 2.1. In this figure, I demonstrate the ordering of rows and columns that I will now use throughout. Columns are ordered by decreasing difficulty so that the most difficult items are at the left. Rows are ordered by increasing ability so that the lowest ability individuals are at the top. The effect of this is to have the cells with the lowest probability of a correct response at the top left and the highest probabilities at the bottom right. Moving on any path downwards and right through the cells, the probability of a correct response must be increasing if the Rasch model holds. However, moving up and right (or down and left) leads to areas of uncertain probabilities. In the top right, you have the probability of correct responses for low ability individuals to easy items. At bottom left, the probability for high ability individuals to hard items. The effect of double cancellation is to ensure enough ordering (for the theorems demonstrating the existence of
interval scalings) on the paths along these minor diagonals. Consider this $3 \times 3$ grid of cells.

$$(1, 1) \quad (1, 2) \quad (1, 3)$$

$$(2, 1) \quad (2, 2) \quad (2, 3)$$

$$(3, 1) \quad (3, 2) \quad (3, 3)$$

For one version of double cancellation to hold, if the value in cell $(2, 1)$ is less than the value in $(1, 2)$ and the value in $(3, 2)$ is less than the value in $(2, 3)$, then it must be the case that the value in $(3, 1)$ is less than the value in $(1, 3)$.

The ability to manipulate the levels of the construct independently of one another is a crucial feature of conjoint measurement. It is important to note that, from a mechanical perspective, Rasch measurement can handle such manipulation since it allows for “specific objectivity” (Rasch, 1960), or the idea that comparisons of individuals or items should be independent of the other. On the item side, the goal would be to create items that are at specified levels of the attribute. This is difficult, but the tradition of item difficulty modeling (where one attempts to understand the difficulty of an item via its various features) suggests it is not impossible. Work along this line has a history dating back to the early 1970s when Scheiblechner (1972) established that item difficulty can be broken down into a combination of parameters related to cognitive processes needed to
obtain the solution. Fischer did similar work establishing the linear logistic test model as a model for detecting change (for an overview of this work, see G. H. Fischer, 1995). The Lexile scale (Stenner, Smith, & Burdick, 1983) perhaps does much of what is required. On this scale, reading comprehension items are hypothesized to be difficult due to a combination of sentence length and word complexity. Extensive work along these lines now allows for the estimation of item difficulty using measures of the two relevant attributes (sentence length and word difficulty, note that one can use “measure” in the classical sense for these two attributes) prior to administering an item to children.

Manipulation on the side of the individual is less developed (in educational assessment, at least). In some sense, it should clearly be possible since the entire goal of education is to manipulate the construct in question. One possible approach might involve delivering “units” of instruction on a given construct. Students would be grouped at common levels based on how many units of instruction they have received. I see two potential problems with this approach. First, the length of the units (1 class? 1 semester?) could be problematic as there would be a trade-off between accuracy on the ordering of individuals and accuracy of the scale. It is fairly easy to order individuals based upon their mathematical knowledge at ages 5, 10, and 15 (where they have received roughly 0, 5, and 10, “units” of mathematics respectively, a “unit” here being one academic year of mathematics instruction). But a scale based upon ordering rows in this way would be quite coarse. On the other hand, finer scales would require orderings based upon smaller units, and this would lead to rougher classifications of the rows. The second problem has to do with a potential conflation between “units” of instruction and ability on the construct. Interest is in assembling an interval scale for the latter, but the grouping of the rows is being based on the former. Individuals learn at very different rates and start at very different places. I am not convinced that this fact can be aligned with the units of curricula approach. This dissertation, however, deals with a different methodological issue and so is not affected by these open questions.
2.5 Are Rasch measurement and ACM compatible?

2.5.1 The role of error

There are some interesting differences between the Guttman (1941) method of scale construction and the Rasch (1960) model. Suppose that we have an individual with ability $X$ and an item of difficulty $Y$. The Guttman model proposed that the individual would answer the item correctly if and only if

$$X > Y.$$  

This scheme offers only purely ordinal characterizations of the attributes involved. Given a large number of responses to an item by an individual, we could only say which items they were “between”. The Rasch model, on the other hand, supposes that an item is answered correctly if

$$X + e > Y$$

where $e$ is an error drawn from the inverse hyperbolic tangent function (for logistic ogives). This difference was noticed by Michell (2008a). He dubbed it the Rasch paradox: the introduction of error is the only thing that separates the two, and yet the latter may allow for interval scale construction. Michell (2004) goes on to make the point that a normal ogive Rasch model would require normally distributed errors. The difference between error distributions have ramifications. If the attributes being investigated are actually quantitative then this choice is important. Since little aside from convenience motivates such a choice, “psychometricians derive what they want most (measures) from what they know least (the shape of error) by presuming to already know it” (p. 126). If one simply admits they do not know the shape of the error distribution, then only ordinal measures can be obtained unless cancellation conditions are checked (Grayson, 1988).

---

12 The other problem with the Guttman model is that it requires a strict ordering of item responses that is rarely met in practice. Wilson (1989) found that roughly 40% of students had problematic responses, from the Guttman ordering viewpoint, on a certain assessment. This is far too large a number for the model to be useful in practice.

13 Andrich (2003) frames this as a trade-off between precision and consistency. Guttman scales are hard to develop in most instances since the orderings of interest are necessarily quite fine. He offers an example of a series of items, ranging from basic arithmetic to calculus, that should be Guttman-like. Such a scale would not be very useful, however. Although Andrich’s point is well taken, I do not see how it addresses the paradox identified by Michell.
What does $e$ mean? Where does it come from? These two seemingly simple questions have inspired decades of debate. Lord and Novick (1968) introduce error through their infamous “brain-washing” thought experiment. They wanted to invoke the theory of errors that had been used to great effect in other scientific fields, but other fields had been able to leverage repeated observations. Repeated observations in testing are unlikely to have the desired properties since later observations are likely to show the influence of fatigue, learning, boredom, etc. To bypass this issue, they introduced the idea of administering a test repeatedly to the same individual, but removing the effects of the subsequent administrations through brain-washing. This idea was formalized in the “stochastic subject” assumption (Holland, 1990). This assumptions suffers from the basic fact that it is difficult to determine what is causing the variability in the responses on each administration. Borsboom (2005, p. 36–38) cites 3 possible sources of variability and finds all of them wanting. An alternative would be to assume that individuals are sampled from some larger population. This “repeated sampling” interpretation (e.g., Meredith, 1993) requires the additional assumption that the measurement model is identical within and between subjects. This assumption, known as “local homogeneity”, is unfortunately not testable and leads Borsboom (2005) to conclude the error structure is ill-defined.

As a contrast to the Rasch model, using the measurement axioms as a completely external frame of reference leads to some simplification of these problems. First, weaker assumptions are being made about the response model and this should lead to a more plausible local homogeneity assumption. More importantly, the grouping of individuals at common levels of the attribute is advantageous since one no longer needs to introduce brain-washing. However, there are novel sources of error in this approach. I illustrate through the following hypothetical scenario. A test is administered to students and the Rasch model is going to be used to scale the responses. (Suppose that pilot testing supports the use of the Rasch model.) Students are grouped into rows representing common ability by collecting all those with the same total score, the sufficient statistic for ability if the Rasch model is true (Baker & Kim, 2004). Rows and columns are then ordered via sum scores. This data matrix could then be subjected to the kinds of analysis I describe in the next chapter.
What kinds of error are introduced in this approach? There are several potential sources.

- Individual entries in the response matrix (that is, responses from an individual to a single item), will suffer from the same problems identified in the Rasch case. However, these entries are going to be aggregated and this calculation of a proportion leaves everything on a surer footing as a theory of errors that does not depend upon brain-washing can now be articulated. Namely, the individuals within a common row are acting as repeated observations.

- However, individuals have to be placed within a common row and here there is an additional possibility of error. What is imperative is that individuals grouped within rows are at the same level of the construct. There are at least two ways this could plausibly go wrong. First, individuals could just be misclassified. For example, if the 3pl model is the true data generating model, then the sum score is not an appropriate means of ordering respondents. Using sum scores will be quite likely to lead to individuals at different true levels grouped into the same rows, or estimated levels. Second, individuals grouped within rows may be at the same level of construct, when grouped with one degree of accuracy, and yet deserve to be broken down into multiple rows when viewed with another level of accuracy. Imagine a general purpose mathematics exam of the type found on the SAT. Suppose using sum scores from such an exam to group students is defensible. If students were then administered an advanced exam on geometry (say consisting of Putnam-exam type questions), the initial ordering might be far too rough to be of much use, especially at the high end of the ability distribution.

- A final source of error is in the ordering of the columns. Columns (representing items) need to be ordered from most to least difficult. Errors could be made in the ordering. When trying to order items on the Lexile scale, for example, Kyngdon (2011) ran into a potential mis-ordering. He found such the error to be relatively minor since the overlap between the two items was within measurement error and reversing the ordering resolved the problem.
As I explain in subsequent Chapters, this is not a solution that will be generally tenable.

Conclusions made on the basis of mis-ordered items would clearly be of dubious value.

What is common to the three named sources is that they lead to errors in the assembly of a data matrix. Such errors will propagate through the subsequent analyses, leading to results that are either not necessarily of interest or potentially misleading. In summary, the role of error in the Rasch model has been found quite tenuous from a philosophical point of view. I think that there is reason to believe some of the most severe critiques of error in the Rasch scenario would not apply to conjoint measurement. On the other hand, the entire concept of error in the construction of a data matrix hasn’t been formalized to the same degree. The main drawback here is that it is difficulty to estimate the effect of different amounts of error.

2.5.2 Philosophical distinctions between RM and ACM

Although previous scholars have noted the similarities between the ACM and the Rasch model, Kyngdon (2008b) disputes this claiming that “from the perspective of RMT [representational measurement theory], the Rasch model is not conjoint measurement” (p. 100). This is due to the reliance on numbers (e.g., probabilities) in the argumentation that the Rasch model is a probabilistic formulation of conjoint measurement. Representationalists begin with an empirical relation system, something that is directly observable. Kyngdon believes probabilities to be an inappropriate choice for the basis of an empirical relation system, and that while the two may be analogous they can not be the same. That said, he does see potential utility in using the methods developed by representationalists for testing claims about psychometric scales. This should sound familiar: “Michell assessed my arguments from a realists perspective of measurement and arrived at a conclusion indistinguishable from the one I drew” (p. 125 Kyngdon, 2008a). Leaving aside the dispute about the nature of number, there are more concrete distinctions that one can draw between the two measurement paradigms.

Karabatsos (2001) suggested item response curves as one way of understanding the difference between Rasch measurement and ACM. Non-crossing item response curves are fundamental to
the ACM theory demonstrating interval scales to be possible. The entire litany of cancellation assumptions (single, double, triple, ...) must be shown to hold before such curves are known to exist. However, the Rasch model imposes non-crossing item response functions on data via the assumption that the functions are of a given form (e.g., logistic). The non-crossing curves can result from either approach (data fitting the form specified by the Rasch model or upholding the ordering assumptions of ACM). One must be wary, however, as item response functions estimated using the Rasch model are by definition non-crossing since they are based on the assumption that the data fits the model. Intervally scaled variables, on the other hand, are only possible when the data are consistent with the axioms. For this reason, Rasch model fit is a natural subject of interest.

Kyngdon (2011) points out a disconnect between the estimated parameters from a Rasch model and the level of the attribute. Assuming that psychological variables are at least ordinal, we can detect the joint effect of certain levels of ability and item difficulty via the observed proportion of individuals at the specified ability level who answered an item at a given difficulty correctly. These observed proportions can certainly be compared to probabilities of a correct response using abilities and difficulties estimated with the Rasch model. Despite the fact that they are seemingly similar, “the set [of person abilities] does not correspond to the set of Rasch person parameters ... as the former comprises the levels of an empirical attribute and the latter is a set of numerical parameters” (Kyngdon, 2011, p. 482). What this hints at is the fact that arguments based on the observed proportions are going to be difficult to extend to arguments about the estimated parameters.

More specifically, if a certain set of axioms is found to be consistent with the structure of data generated as a byproduct of two factors, then an interval scale for the factors is guaranteed to exist but this is not an assurance that the scale created by a certain IRT model is interval. This difference leads Kyngdon to suggest that we are looking for “plausible analogies” between axiomatic measurement and psychometric models. At first blush, it may seem disappointing that the methods used in this dissertation will only demonstrate whether the scales are plausibly interval. I would
argue that this is a substantial improvement over merely assuming that they are. To see this, consider again the two possibilities on the right hand side of Table 1.1. Knowing that the data from one assessment is in cell II means that interval scales are possible. Additional tools from the Rasch modeling toolkit, such as those that deal with multidimensionality or local dependence, may lead to a defensibly interval scale. Data that is in cell III, on the other hand, will not lead to interval scales no matter how many additional statistical techniques are applied.\footnote{This may be an overstatement. It could be shown that (A) some alternative model fits the data and that (B) this model can lead to interval scales. Although the model required in (A) may be simply a more complex item response model, there is no good candidate for an alternative argument for interval scales in (B). It seems safe to simply presume that interval scales are impossible.} What would be useful, statistically, for data in this quadrant are the range of techniques available for ordinal measures (e.g., Johnson & Albert, 1999; Agresti, 2010), assuming that ordinal scales are even feasible.

Before moving forward, it is worth noting that not all scholars believe the advanced made in representational measurement to be so promising in the quest for interval scales for psychological variables. While Cliff (1992) called axiomatic measurement “the revolution that never happened” in psychometrics, Trendler (2009) suggests that it is a revolution that “cannot happen”. Trendler suggests that psychology will never be able to construct scales in a manner consistent with representational arguments since the nature of the mind allows for limited control and manipulation. Although this claim may one day be shown to be true, I think that there may still be worthwhile insights to obtain from using the two methods, ACM and the Rasch model, in tandem.
Chapter 3

Methodology

There is no question that interval scales for psychological variables are greatly desired. Such scales would be much easier to study analytically and would make communication about the scale much simpler. This desire seems to frequently lead people to assume that scales are interval: Harwell and Gatti (2001) conducted a survey of three journals (American Educational Research Journal, Sociology of Education, and Journal of Educational Psychology) in 1997 and found that 73% of the dependent variables mentioned in the contained articles only appeared to justify ordinal interpretations. However a large percentage of studies in these journals used techniques (such as hierarchical linear modeling) that are only appropriate, according to the Stevens (1946) typology, with intervally scaled outcome variables. Such an assumption could lead to erroneous conclusions since the scales are assumed to have properties that they do not possess.

The axioms of conjoint measurement are one way of demonstrating that interval scales are possible for psychological variables. This chapter discusses articles relevant to the application of the concepts of conjoint measurement to educational assessment. After discussing this body of work, the specific methodological approach advanced by Karabatsos (2001) and extended by Kyngdon (2011) is described in detail. After noting some shortcomings with these studies, I describe an updated methodology. In closing, the item fit statistics that are used in subsequent chapters are presented.
3.1 Literature Review

As Harwell and Gatti (2001) demonstrate, researchers frequently assume that scales are equal-interval despite the knowledge that this assumption may be unwarranted. In a rather telling twist, after Harwell and Gatti discuss the dangers of mistakenly assuming that ordinal scales have the equal-interval property, they then make the claim that “under some conditions described later, IRT produces interval-scaled data” (p. 109). As Michell (2008b) has emphasized, there is nothing in the theory of item response modeling that guarantees interval scales merely from fitting the model. Unfortunately, the “conditions” that Harwell and Gatti describe seem typical of many non-critical application of IRT models. For example, they justify their choice of a Rasch model with dichotomously scored data by saying “we had no reason to believe that the items varied in discrimination or that guessing needed to be modeled” (p. 119). That is, no a priori reason. Their analysis of fit statistics was cursory and yet they still found 10 items (out of 30) that were inadequate. While they are correct about the need for distinctions between interval and ordinal scales, there is no reason to uncritically assume that the estimation of parameters via IRT models is going to yield interval scales.

The remainder of this section will focus on the relatively small body of research describing methods for checking whether assessment data may meet some form of axiomatic conjoint measurement. There is one category of examples I will discuss only briefly as it is requires comment but is not my focus. The Rasch community (distinguishing it from the sometimes overlapping IRT community) has long been aware that the Rasch model is related to ACM (Keats, 1967; G. Fischer, 1968; Brogden, 1977; Karabatsos, 2000). Many papers have attempted to leverage the conjoint measurement machinery without showing that the axioms hold, but by showing instead that a certain Rasch model fits. The special issue introduced by Fisher Jr and Wright (1994) contains several examples of the genre. They note that “the purpose of the following chapters is to describe applications of probabilistic formulations of conjoint measurement in various areas of the behavioral sciences” (p. 559). However, most of the application they describe are relatively
straightforward uses of the Rasch model. G. N. Masters, Adams, and Lokan (1994), for example, describe a system they have developed for the communication of student achievement.\footnote{I find it intriguing that this system communicates information from the supposedly interval scale in a primarily ordinal way, see their Figure 3.1.} This system is based on creating a scale using estimates from the partial credit model (G. Masters, 1982) and, presumably, relies upon fit statistics to reach a conclusion that the model fits and, therefore, probabilistic conjoint measurement is possible. No reference is made to the underlying axioms of conjoint measurement. Esdaille, Shaw, Smith, and Valgeirsdottir (1994) proceed similarly. These articles are related to this study in that each is aware of the role of conjoint measurement if the resulting scale is to be treated as interval. In their approach, the Rasch model is invoked as “probabilistic” conjoint measurement and interval scales are assumed to be a byproduct of data fitting the Rasch model. In my approach, I attempt to directly answer whether data supports the axioms of ACM and, only then, are questions related to model fit relevant. My primary interest in this review are those studies which have been specifically focused on scale properties as they relate to conjoint measurement, perhaps alongside the Rasch model as a scaling mechanism. An early example is Perline, Wright, and Wainer (1979). This paper demonstrates the correspondence between the two viewpoints and analyzes empirical data to determine whether the axioms seem to be consistent with the data. Two aspects of this research are important as they are similar in spirit to what I will do.\footnote{Another major component of their work that I do not focus on was comparing estimates from the Rasch model to those obtained via another methodology (Kruskal, 1964) specifically designed to yield additivity. They found that the Rasch model performed well in comparison to the alternative approach.} First, the authors focus on Rasch fit statistics as one indication that the Rasch model hold but they also examine the axioms directly. Second, they test all possible 3-matrices for double cancellation, something that has not been widely attempted in later work. This was possible in part due to the relatively small set of items that they chose to work with. They make use of two data sets. In the first, they find violations of cancellation axioms in roughly 10% of the tested cases. The second data set demonstrates stronger support for the axioms, under 5% of their cancellation tests show violations. These authors clearly believe the idea of the Rasch model as an instantiation of conjoint measurement to be a promising idea.
One limitation of the Perline et al. (1979) study was that it did not allow for measurement error. The observed proportion of correct responses (used to check the cancellation axioms) were treated as fixed quantities observed without error. Falmagne (1976) developed methods that would allow one to introduce measurement error into the testing of the axioms of ACM. Falmagne did this in an attempt to construct a scale for loudness. That method relied strongly upon an experimental setting that is perhaps not yet possible for educational assessments. This initial work was extended by Iverson and Falmagne (1985) and Davis-Stober (2009). Together, these articles offer a frequentist alternative that could be pursued, in theory, as an alternative to the Bayesian methodology employed here.

Green (1986) was influenced by the work of Iverson and Falmagne (1985) in that the author attempts to develop a methodology for evaluating the axioms with certain data thought to be measured with error. The method is relatively straightforward. Confidence intervals for the probability of observing a correct response in each cell of a conjoint matrix are computed and then used to test whether differences between cells hold at some level chosen by the researcher. Using this technique, the author determined that single and double cancellation were both violated using data generated from 19 multiple-choice (general knowledge) items given to over 900 undergraduates. This method seems to have relatively low power, and I wonder if it is not related to the following observation: “in this application independence and double cancellation were null hypotheses” (p. 146). That is, the truth of these axioms was being evaluated post-hoc. In contrast, the methodology I employ to test the axioms is based on using the axioms to establish the confidence intervals for relevant quantities. The specificity in the approach I use comes at a loss of generality. Green’s technique could easily be used to test, for example, the third cancellation axiom since it doesn’t require the extensive manipulation discussed later in this chapter.\footnote{The author does do something rather interesting: “Multiple-choice test items were systematically varied along two dimensions (language difficulty and semantic similarity of response options) and the effects on item difficulty assessed” (p. 141 Green, 1986). Such systematic manipulation of item characteristic in an attempt to control item difficulty is relatively rare and similar to the manipulation done as part of the Lexile scale.}

Karabatsos (2001) devised a Bayesian method for checking probabilistic version of the axioms.
His work forms the basis for the methodology that will be used in this study. The key idea is to enforce the axioms stochastically and then determine whether the resulting nonparametric estimates for the probability of a correct response are reasonable given the observed data. I discuss this methodology in the next section. In this paper, he re-analyzes the data originally used by Perline et al. (1979) but now allowing for measurement error. His analyses, however, are rather cursory (for instance, he only looks at a single instance of double cancellation compared to the complete approach in the original work), so I don’t think much is added to the understanding of the original data.\footnote{His analysis of a single $3 \times 3$ matrix is interesting, but less useful than the extensive analysis of such matrices conducted in Perline et al. (1979) that demonstrate violation in nearly 10\% of all cases.}

Kyngdon (2011) builds upon this work by applying the methodology to a dataset far closer to the types seen in large-scale educational assessments than that used by Karabatsos (2001) and also by suggesting a philosophical stance that is more likely to be embraced by practitioners. To contrast the philosophies, let us begin with Karabatsos’s description of his work:

This research introduces non-parametric item response models, based on ideas of order restricted inference (Robertson, Wright, and Dykstra, 1988), that estimate IRFs and PRFs under the null hypothesis that the data accord with a given set of measurement axioms. (p. 397)

In contrast, Kyngdon views particular IRT models as possible analogies of conjoint measurement. This is motivated by his distinction between the variables one manipulates in conjoint measurement, which are (hypothesized) levels of an empirical attribute, and the estimates from IRT, which are simple numerical parameters (2011, p. 482). The non-parameteric models suggested in Karabatsos’s work are unlikely to be used by large-scale test developers given their extensive use at present of parametric IRT models. Given this reality, it makes more sense to encourage states and contractors to consider the salient axioms of conjoint measurement without requiring they leave their existing framework entirely. In doing so, they will be forced to conduct additional analyses related to the strength of their scale rather than relying on fit statistic alone. This will bolster the credibility of their scale and help to validate it for uses that require interval properties (such as some value-added
models.\textsuperscript{5}

Kyngdon’s (2011) study applied the axioms of conjoint measurement to data generated from the Lexile (Stenner et al., 1983) assessment. This assessment has a number of attractive characteristics from the perspective of ACM. He tentatively stated that “the difficulty of reading items, as conceived of in the Lexile theory, and the reading ability of persons, are quantitative” (Kyngdon, 2011, p. 488). This was based on checking single cancellation and the two instances of double cancellation from within a single $3 \times 4$ conjoint matrix. Since there are many abilities not represented and more items that go into constructing the scale, this would at first glance seem to be an insufficient test of the quantitative hypothesis. However, Kyngdon argued that previous research (McClelland, 1977) demonstrated that a randomly chosen matrix was extremely unlikely to support the axioms by chance if the axioms were not supported in the data matrix as a whole. He was able to verify that the cancellation axioms held in his single $3 \times 4$ matrix, but only after reversing the ordering of two items. That is, the difficulties of the items establish by the Lexile theory were not consistent with single cancellation but the axioms did seem supported when he reversed the ordering.\textsuperscript{6} I will argue that his finding of a quantitative structure is premature.

\subsection*{3.2 Karabatsos’s Methodology}

The methodology used in this dissertation originates in Karabatsos (2001).\textsuperscript{7} Suppose that the responses to a set of items by a set of persons are functions of only a single characteristic of the item and a single characteristic of the individual. Moreover, imagine that items and individuals can be ordered along this characteristic. If Karabatsos’s stochastic versions of the ACM axioms hold over the Cartesian product of these sets, then a non-interactive function of the two characteristics will produce the probabilities of a true response to an item by a person. Let $P$ be the $J \times N$ matrix

\textsuperscript{5} It is worth re-emphasizing that the consumers of data derived from educational assessments have been perhaps overly credulous since claims such as those found in CTB—McGraw Hill (2001), claiming that the TerraNova has equal-interval properties, have gone largely undisputed.

\textsuperscript{6} The distance between the two items’ established difficulties was within the known error of measurement for the Lexile difficulties, so this was not an especially egregious reversal.

\textsuperscript{7} Karabatsos develops much of his work using the models from Scheiblechner (1995, 1999). However, I see no reason to do so. These models are not necessary and merely introduce additional notational and theoretically complexity. In this way, as in several others, my presentation is more along the lines of Kyngdon (2011).
that contains the true response probability. Each cell in the data matrix, denote it $\hat{P}^{\text{MLE}}$ with dimensions $J \times N$, contains the percentage of respondents with a certain ability who answered the appropriate item correctly. Obviously $\hat{P}^{\text{MLE}}$ is an estimate of $P$.\(^8\) Due to measurement error, $\hat{P}^{\text{MLE}}(j,n)$ is not necessarily $P(j,n)$.

The novel idea behind Karabatsos’s work is that the order restrictions inherent in the axioms are imposed stochastically. This idea was first suggested by Devroye (1986) and adapted to the MCMC context in Gelfand, Smith, and Lee (1992). I begin with a brief example involving single cancellation. Concentrate on $P(j,n)$ with $1 \leq j \leq J$ and $1 \leq n \leq N$. Under the same assumptions on the “direction” of row ordering that is shown in Figure 3.1, independence implies that $P(j-1,n) < P(j,n) < P(j+1,n)$ for each $n$. We use a Metropolis-Hastings (MH; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970) algorithm to find an alternate estimate for $P(j,n)$. Consider only column $n$. Begin with a matrix of initial estimates, $\hat{P}^0$ that are ordered appropriately. At iteration $t$, the proposal distribution for cell $(j,n)$ is

$$\text{Unif} [\hat{P}^t_{(j-1,n)}, \hat{P}^t_{(j+1,n)}] \quad (3.1)$$

\(^8\) This estimator is in fact the maximum likelihood estimator for $P$ since we are not imposing any response model outside of the fact that individual responses are Bernoulli random variables.
if one abuses notation and fixes $\hat{P}_{(0,n)}^t = 0$ and $\hat{P}_{(j+1,n)}^t = 1 \forall t$. What is of fundamental importance to note is that the ordering implied by the axiom of independence is clearly being imposed via the proposal distribution since the choice of proposal implies that $\hat{P}_{(j,n)}^t$ will be in the interval $[\hat{P}_{(j-1,n)}^t, \hat{P}_{(j+1,n)}^{t-1}]$. A sample is drawn from the proposal and then is accepted or rejected via the standard MH acceptance ratio (discussed in the next section). After burn-in and thinning, a 95% credible region is formed from the remaining $\hat{P}_{(j,n)}^t$. If each $\hat{P}_{(j,n)}^{MLE}$ falls within the appropriate credible region, then the axiom is said to hold stochastically. If not, then violations can be pinpointed to specific cells. For example, in Table 3 of Karabatsos (2001), cell (3,1) has a posterior with 95% credible region spanning from 0.57 to 0.725. In fact, 60 out of 82 respondents answered this item correctly so the MLE estimate would be 73.2%. Since $0.732 \notin [0.57, 0.725]$, a violation (though slight) is detected.

Checking double cancellation is more difficult. Going back to Figure 3.1, the smallest probabilities of a correct response are found at upper left and the largest at bottom right. However, as you move away from this main diagonal at right angles, it is unclear exactly how the probabilities are changing as in one direction the items are getting harder but the individuals more able and in the other the items are getting easier but the individuals less able. As Karabatsos (2001, p. 414) notes, extending this method to the higher order cancellation axioms (above single cancellation) is challenging since these checks depend on the selection of a sub-matrix from the data matrix (a 3-matrix in the case of double cancellation). An additional complication seems to be that each 3-matrix could present numerous versions of double cancellation. Writing double cancellation generically as

$$f(x_i, y_j) \geq f(x_j, y_i) \& f(x_j, y_k) \geq f(x_k, y_j) \rightarrow f(x_i, y_k) \geq f(x_k, y_i),$$

then there are 36 possible examples of double cancellation one can obtain by changing the values of the indices. There is the version we have already encountered (see section 2.3):

$$f(x_1, y_2) \geq f(x_2, y_3) \& f(x_2, y_3) \geq f(x_3, y_2) \rightarrow f(x_1, y_3) \geq f(x_3, y_1)$$
where \( i = 1, j = 2, \) and \( k = 3. \) Choosing instead \( i = 2, j = 3, \) and \( k = 1, \) we get
\[
f(x_2, y_3) \geq f(x_3, y_1) \& f(x_3, y_1) \geq f(x_1, y_3) \rightarrow f(x_2, y_1) \geq f(x_1, y_2).
\]
And so on. Michell (1988) showed that these 36 instances reduce to six clusters where the truth value of all instances within each cluster are identical (see Figure 4.3 in Michell (1990) for a schematic). Furthermore, if single cancellation holds, then double cancellation will hold in five of the clusters. So within a 3-matrix, given single cancellation we may speak unambiguously of the test of double cancellation\(^9\) (which I take to be the version given when \( i = 1, j = 2, \) and \( k = 3).\)

### 3.2.1 The Metropolis-Hastings algorithm

For any given conjoint matrix, I will index the items, or columns, by \( j \in \{1, \ldots, J\} \) and the collections of respondees (remember that in a conjoint matrix respondents have already been aggregated into rows on the basis of their assumed common ability), or rows, by \( i \in \{1, \ldots, I\}.\)

Let \( Y_{ij} \) be the binomial random variable\(^{12}\) associated with answering an item correctly. The goal is to understand whether the probabilities underlying these random variables (assuming that the other parameter for this binomial random variable, the number of trials, has been fixed), denoted as \( P_{ij}, \) are consistent with the axioms of additive conjoint measurement. This is done by first obtaining a posterior estimate for the parameter using a Metropolis-Hastings algorithm (MH; Metropolis et al., 1953; Hastings, 1970) where the ordering restrictions implied by the measurement axioms are imposed through the jumping distribution and, secondly, assessing whether the observed proportion of correct responses is consistent with this posterior.

---

\(^9\) This is actually a mis-reading of the point made in Michell (1990). Clarification can be found in the Appendix.

\(^{10}\) This is, perhaps, a slight over-simplification. If single cancellation holds, then there is indeed a single test of double cancellation necessary. However, in my analyses, I am testing both single and double cancellation at once. Since single cancellation is not known to hold, is it not premature to test for only a single instance of double cancellation? I think, in fact, that it is not premature. I believe that my proposed method is, for a given 3-matrix, the more stringent of the two tests. Single cancellation must always hold and additional requirements for double cancellation (assuming single cancellation) are added. So if there are no detected violations, then single and double cancellation both seem to hold. If there are detected violations, it is not immediately apparent whether single or double cancellation is at fault, but one would not be led astray (that is, one would at least know there is a violation of some kind).

\(^{11}\) This notation differs from that normally observed in IRT (where \( i \) typically indexes columns since they represent items), but I think it is preferable here to emphasize the matrix notation.

\(^{12}\) Although it is no longer a Bernoulli random variable due to the aggregation by row, the \( P_{ij} \) is still the probability of a student in row \( i \) answering item \( j \) correctly.
Without imposing any model for an item response, the maximum likelihood estimate for the probability of a correct response for cell \((i, j)\) is simply the proportion of respondents at ability level \(i\) who answered item \(j\) correctly. If \(n_{ij}\) respondents answered correctly out of \(N_{ij}\) respondents, then 
\[
\hat{P}_{i,j}^{\text{MLE}} = \frac{n_{ij}}{N_{ij}}.
\]
Consider now the joint likelihood. Under standard independence assumptions, we have
\[
L(P|N, n) = \prod_{i=1}^{I} \prod_{j=1}^{J} \binom{N_{ij}}{n_{ij}} P_{ij}^{n_{ij}} (1 - P_{ij})^{N_{ij} - n_{ij}} \tag{3.2}
\]
where non sub-scripted quantities are shorthand for vectors (e.g., \(P = (P_{11}, ..., P_{IJ})'\)). I assume throughout that everything is ordered in the standard way (rows/people from least to most able going top to bottom and columns/items going from hardest to easiest from left to right) such that only the single test of double cancellation is necessary.

### 3.2.2 Single Cancellation

Single cancellation requires that both:

- \(\forall j P_{i_1,j} < P_{i_2,j}\) so long as \(i_1 < i_2\).
- \(\forall i P_{i,j_1} < P_{i,j_2}\) so long as \(j_1 < j_2\).

Since the two conditions are so closely related, I will only work with the second. Beginning with a set of initial estimates for each \(P_{ij}\), call them \(\hat{P}_{ij}^0\), create a MCMC chain of estimates in each cell, starting with cell \((1, 1)\). After accepting or rejecting a draw from the jumping distribution, the \(t = 0\) estimate in cell \((1, 1)\) is updated to \(\hat{P}_{11}^1\). This updated estimate, along with the initial \(t = 0\) estimates for all other cells, is then used to update cell \((1, 2)\). This process repeats until the \((1, J)\) cell is updated, at which point one begins again in the second row. Note that when considering the \(t\)-th iteration for cell \((i, j)\), cell \((i, j - 1)\) will have estimate \(\hat{P}_{i,j-1}^t\) while cell \((i, j + 1)\) will have estimate \(\hat{P}_{i,j+1}^{t-1}\).

We obtain the \(t\)-th estimate for \(P_{ij}\) as follows. Use the jumping distribution
\[
Q_{ij} = \text{Unif}[\hat{P}_{i,j-1}^t, \hat{P}_{i,j+1}^{t-1}] \tag{3.3}
\]
For the first and last rows in the ordered conjoint matrix, we set the lower, $\hat{P}_{0j}$, and upper, $\hat{P}_{I+1,j}$, bounds of the jumping distributions to 0 and 1 respectively. I apply the same approach to the first and last items (columns). Draw $\hat{P}_{ij}^*$ from this distribution. The acceptance ratio is

$$ r = \frac{Pr(\hat{P}_{ij}^*)Q_{ij}(\hat{P}_{ij}^*|\hat{P}_{ij}^{t-1})}{Pr(\hat{P}_{ij}^{t-1})Q_{ij}(\hat{P}_{ij}^{t-1}|\hat{P}_{ij}^*)}. $$

(3.4)

Draw $t$ from $\text{Unif}[0,1]$. If $r > t$, then set $\hat{P}_{ij}^t = \hat{P}_{ij}^*$. Otherwise, set $\hat{P}_{ij}^t = \hat{P}_{ij}^{t-1}$. The acceptance ratio can be greatly simplified. First, the particular choice for jumping distribution is such that $Q_{ij}(\hat{P}_{ij}^*|\hat{P}_{ij}^{t-1}) = Q_{ij}(\hat{P}_{ij}^{t-1}|\hat{P}_{ij}^*)$ (since, in fact, neither of these terms appear in Eqn 3.3) and so those terms cancel. Second, if we assume a flat prior, then the $Pr(\hat{P}_{ij}) = L(\hat{P}_{ij})$. Moreover, the double product in Eqn 3.2 is unnecessary since all terms involving cells other than $(i,j)$ cancel. Finally, the binomial coefficient cancels and we are left with

$$ r = \frac{(\hat{P}_{ij}^*)^{n_{ij}}(1 - \hat{P}_{ij}^*)^{N_{ij} - n_{ij}}}{(\hat{P}_{ij}^{t-1})^{n_{ij}}(1 - \hat{P}_{ij}^{t-1})^{N_{ij} - n_{ij}}}. $$

(3.5)

In practice, one actually works with the log-likelihoods to avoid computational singularities.

To understand what is happening, consider the case when $n_{ij} \to 0$. We have

$$ r \to \left( \frac{1 - \hat{P}_{ij}^*}{1 - \hat{P}_{ij}^{t-1}} \right)^{N_{ij}}. $$

If $\hat{P}_{ij}^*$ is large relative to $\hat{P}_{ij}^{t-1}$, then $r$ will be very nearly 0. Since $n_{ij}$ is small, the probability of a correct answer must be small and the algorithm would prefer to stick with $\hat{P}_{ij}^{t-1}$ in this case, rather than update to the much larger $\hat{P}_{ij}^*$, since the latter would not lead to a low $n_{ij}$.

Chains are formed using this approach for all cells. After convergence to the posterior distribution has been reached, each chain of estimates is burned and thinned with the resulting values being treated as a set of independent draws from the posterior distribution. In particular, we can use these draws to estimate the 95% credible region for the parameter. The data supports the single cancellation axiom being tested here if $\hat{P}_{ij}^{\text{MLE}}$ lies in this 95% credible region. The logic is that if the observed proportion is not in the credible region, then the true response probability for the cell is unlikely to conform to the measurement axioms imposed via the jumping distribution.
Before moving forward, note a characteristic of the check for single cancellation. Unlike double cancellation, where a 3-matrix is needed, this check could be done in an entire data matrix. That is, the entire data matrix could be checked in one big iteration of the algorithm described above. Indeed, my description of the approach is general enough to allow it. Karabatsos (2001) and Kyngdon (2011) both used these checks in small matrices (3 × 3 for the former and 3 × 4 for the latter), so such an observation would not have made much difference. On the one hand, if only a single check of single cancellation needed to be performed, this would be much more efficient. On the other hand, note that if one were to attempt this approach over an entire data matrix, the resultant proposal used in Karabatsos’s method depends on the number of rows or columns. The larger the number, the finer the proposal “mesh” that is formed. As the mesh gets finer, the MLE estimates are going to have to fit within tighter credible regions. In contrast, it is no more difficult to fit the Rasch model to a larger group of items than a smaller one (the emphasis here is on the mechanical act of estimation; it is actually the case that an instrument with more items will have a harder time actually fitting the Rasch model), but this is due to the structure endowed by the assumed mathematical form. Since all curves in the Rasch model have the same shape (typically based on the logistic distribution) and only differ in terms of where they are centered, any two items will lead to non-crossing curves if there is any difference in terms of their estimated center. Without that mathematical structure, larger conjoint matrices may be quite unlikely to meet the axioms in this stochastic setting. In initial simulations, I verified that the checks are indeed sensitive to the mesh as hypothesized and thus focused on sampling large numbers of 3-matrices since these do not induce this problem and are also the natural objects to check for double cancellation.

3.2.3 Double Cancellation

Each possible 3-matrix pulled from a conjoint matrix implies a different test of double cancellation. Once a 3-matrix is selected, the check for double cancellation can be described. The only
change from above is the jumping distribution. It becomes

\[
\text{Unif}\left[\max\{\hat{P}_t(i-1,j), \hat{P}_t(i,j-1), \hat{P}_t(i+1,j-1)\}, \min\{\hat{P}_t-1(i+1,j), \hat{P}_t-1(i,j+1), \hat{P}_t(i-1,j+1)\}\right].
\]

(3.6)

It is easier to see via a matrix:

\[
\begin{array}{cccc}
\cdot & \downarrow & \Theta & \uparrow \\
\downarrow & \hat{P}_t(i-1,j-1) & \Theta & \uparrow \hat{P}_t(i,j+1) \\
\downarrow & \hat{P}_t(i+1,j-1) & \uparrow \hat{P}_t-1(i+1,j) & \cdot \\
\end{array}
\]

Here Θ represents the cell whose estimate is being updated. An up arrow represents a cell whose value is greater than Θ since it is in the minimum being taking to determine the right hand edge of the uniform jumping distribution. Similarly, the down arrow represents a cell whose value is less than Θ. Note that in this formulation, one is now ensuring that single cancellation over both rows and columns hold. Kyngdon (2011) seems to have been the first to utilize this fact and design omnibus tests for both types of single cancellation at once. I will adopt this approach as well.

### 3.2.4 A brief semantic digression

I want to provide some detail on the terminology used here because there is a certain amount of imprecision in how certain terms are routinely treated in practice. There are three quantities that one needs be concerned with when applying the Metropolis-Hastings algorithm: the likelihood, the prior, and the jumping distribution (this distribution is also called the proposal or transition distribution). The confusion is between the latter two. Karabatsos (2001) defines the likelihood as what is shown in his Eqn 5 (p. 406),\(^{13}\)

\[
L(n_k|\theta_k) = \prod_{i=1}^{I} \prod_{j=1}^{J} \theta_{ijk}^{n_{ijk}} (1 - \theta_{ijk})^{N_{ijk} - n_{ijk}}.
\]

He defines the prior as what is shown in his Eqn 6:

\[
\pi(\theta_{ijk}) = \begin{cases} 
1 & \min(\theta_{ijk}) \leq \theta_{ijk} \leq \max(\theta_{ijk}) \\
0 & \text{otherwise}
\end{cases}
\]

\(^{14}\) Note that he actually states the likelihood incorrectly since he drops the binomial coefficient. It doesn’t affect his work since this cancels from the acceptance ratio. I reproduce his original version.
The problem is that he obscures the distinction between the jumping distribution and the prior: “a candidate value $\theta_{ijk}^*$ is sampled from the order restrictions defined by the uniform prior distribution” (p. 407). In standard descriptions of the MH algorithm, the candidate value is drawn from the jumping distribution (e.g. Jackman, 2009; Gelman, Carlin, Stern, & Rubin, 2004; Cameron & Trivedi, 2005). However, these textbooks fail to emphasize the importance (or even presence of a prior) in the MH algorithm. For example, in the example given for the MH algorithm in Gelman et al. (2004, p. 290), no prior is ever mentioned and the acceptance ratio is clearly a ratio of only the likelihoods. You can see a similar problem in Example 6.1 of Robert and Casella (2010).

I think this confusion can be partially explained by the following reasons:

• There is a subtle distinction between the Metropolis and Metropolis-Hastings algorithm. The latter differs from the former in that the jumping distribution is not necessarily symmetric. While the acceptance ratio of the Metropolis algorithm is merely a ratio of the posteriors (likelihood times the prior), the acceptance ration in the MH algorithm requires an additional factor that is a ratio of densities for the jumping distributions. Since the jumping distribution is symmetric here (in fact, the jumping distribution is flat), we are actually using the Metropolis algorithm.

• There seems to be a tendency to conflate the jumping distribution and the prior. This is presumably due to the fact that a form of the jumping distribution may be a natural candidate for the prior. However, the jumping distribution can not be the prior since the former is changing at each iteration of the algorithm while the latter must be fixed as it does not depend on the data.

• One possible explanation would be that it is standard practice to assume a flat prior for most parameters and then subsequently ignore them in the computation of the acceptance ratio. There are occasions, however, when priors could be useful. Parameters that we might expect to be only weakly estimable (e.g., guessing parameters in an IRT setting) might benefit from having a prior so that they do not misbehave.
I assume flat priors for all parameters. I have included this discussion to indicate why I have not used the language of Karabatsos (2001) in my description of the MH algorithm and also to indicate that my chosen terminology is consistent with how this method is described in standard texts.

### 3.3 Shortcomings of the original approach

#### 3.3.1 Order along the minor diagonal

Consider the probabilities of a correct response from the Rasch model. If the individuals are ordered from least able (top) to most able (bottom) and the items are ordered from least difficult (left) to most difficult (right), there is a very definite order along the major diagonal. If $P_{ij}$ refers to a true underlying probability for individual $i$ and item $j$, then it will always be the case that $P_{ij} < P_{i+1,j+1}$. There is no such simple statement that can be made about the minor diagonal. The double cancellation axiom exists to ensure some reasonable degree of behavior, but it may help to see just how poorly behaved the minor diagonal can be before moving forward.

I will discuss a few examples, and so that the interested reader may verify my claims I shall include R source (R Development Core Team, 2010).

```r
set.seed(1234321)
5->n1; 10->n2
sort(rnorm(5),decreasing=TRUE)->ip
sort(rnorm(10))->ab
matrix(ab,n2,n1,byrow=FALSE)-matrix(ip,n2,n1,byrow=TRUE)->kern
exp(kern)/(1+exp(kern))->pv
```

It is easy to verify the order in $pv$ moving from upper-left to bottom-right. Let me emphasize a few things about order in the other direction using a 3-matrix (chosen to illustrate the following
points) formed by columns 2 through 4 and rows 4 through 6 of the data simulated in $\mathbb{R}$:

\[
\begin{pmatrix}
0.13 & 0.17 & 0.20 \\
0.19 & 0.25 & 0.28 \\
0.20 & 0.26 & 0.30
\end{pmatrix}
\] (3.7)

Note first that there is no definite order along the minor diagonal here. Clearly $P_{31} < P_{22}$ and $P_{22} > P_{13}$. In fact, $P_{31} > P_{13}$ when more decimal places are used. Does this imply that double cancellation is violated? No. First, let me emphasize that Michell (1990) indicates that only a single instance of double cancellation needs to be tested in this situation. The single instance is that if $P_{21} < P_{12}$ and $P_{32} < P_{23}$ then $P_{31} < P_{13}$. In this example, the antecedent is false since $P_{21} > P_{12}$ (that is 0.19 > 0.17). Therefore the consequent need not hold. It is crucial that one keep in mind that the minor diagonal can, in general, be quite poorly ordered. The relatively weak requirements of double cancellation are the only necessary orderings.

In contrast, if one chooses abilities and item parameters in even intervals:

```r
ip<-seq(1.5,-1.5,length.out=5)
ab<-seq(-1.5,1.5,length.out=10)
```

and then re-computes probabilities, minor diagonals are perfectly ordered. But in this case we are leveraging the interval structure. With empirical data, no such interval structure underlies the row or column orderings and one must work carefully to determine whether double cancellation holds or even needs to hold.

### 3.3.2 Are Karabatsos’s ordering restrictions the right ones?

When checking for double cancellation within a 3-matrix, Karabatsos’s (2001, p. 414) creates a jumping distribution for cell $(i,j)$:

\[
\text{Unif} \left[ \max\{\theta_{(i-1)j}^t, \theta_{i(j-1)}^t, \theta_{(i+1)(j-1)}^{t-1}\}, \min\{\theta_{(i+1)j}^{t-1}, \theta_{i(j+1)}^{t-1}, \theta_{(i-1)(j+1)}^t\} \right].
\]

For simplicity of notation, suppose $i = 2$ and $j = 2$. Using this notation, Karabatsos’s jumping distribution is going to have the effect of ensuring that $P_{31} < P_{22}$ (due to the last value in the
and that $P_{22} < P_{13}$ (due to the last value in the min). Let me illustrate through the data given in 3.7. If those are the present values in the Markov chain and we want to form the jumping distribution for the $(3, 1)$ cell (current value of 0.20), the jumping distribution would be

$$\text{Unif} \left[ \max\{0.19, 0, 0\}, \min\{1, 0.26, 0.25\} \right] = \text{Unif} \left[ 0.19, 0.25 \right].$$

Clearly, the proposed value will be less than or equal to the value in the middle cell, 0.25.\(^{14}\) This is done to ensure that the consequent of double cancellation, $P_{31} < P_{13}$, for this 3-matrix holds. There are two problems here. First, the statement of double cancellation is a conditional. We only require $P_{31} < P_{13}$ if $P_{21} < P_{12}$ and $P_{32} < P_{23}$. Second, we are ensuring that $P_{31} < P_{13}$ through an additional restriction on $P_{22}$. In particular, we are forcing $P_{31} < P_{22} < P_{13}$ which is not only not essential, but also not even necessarily true (as in the example from the last section). Adding these non-necessary requirements makes detection of a distortion more likely but does nothing in terms of ensuring interval properties since the requirement is not motivated by ACM.

Even in the context of using the checks with simulated data, it is basically impossible to know if even the antecedent of double cancellation holds within any given 3-matrix and, consequently, whether the consequent needs to hold. In the simulated data, our knowledge of the underlying probabilities is tied to our knowledge of the true abilities. However, when response data simulated based on those true probabilities are then aggregated, the rows are now groupings via sum score. Since sum scores are not perfectly mapped to the true underlying abilities, we don’t know the underlying true probability for a given cell (perhaps it would be more correct to say that there is none) and thus can’t know the truth value of statements about the ordering of cells. This is a dilemma that I attempt to resolve in the subsequent section using a probabilistic check of the truth of the antecedent of double cancellation before imposing that the consequent be true via the jumping distribution.

The problem of imposing $P_{31} < P_{13}$ through $P_{22}$ is potentially serious\(^ {15}\) but seems fairly

\(^{14}\) This example is not even sufficiently well ordered to use Karabatsos’s approach since the value in the $(1, 3)$ cell is less than the value in the $(2, 2)$ cell. The values in this matrix would need to be substantially re-ordered prior to using Karabatsos’s methodology. My approach does not require as much work in terms of re-ordering.

\(^{15}\) In one simulation with data generated from the Rasch model, fewer than 10% of the 3-matrices I drew to check for
simple to solve. It seems reasonable just to impose the constraints that are actually required:

- For $P_{31}$, let the jumping distribution be $\text{Unif}[\max(P_{21}), \min(P_{32}, P_{13})]$.

- For $P_{13}$, use $\text{Unif}[\max(P_{12}, P_{31}), \min(P_{23})]$.

- For all others, use the constraints that you would for single cancellation.

I am using the fact that work is only occurring in a 3-matrix implicitly here by focusing on the “corner” cells, (1, 3) and (3, 1). Due to the fact that the minor diagonal is so poorly behaved in general, a few additional safeguards will be needed. In the subsequent section, I describe the new approach in full detail and, in the subsequent chapter, give information on its behavior in simulated settings.

### 3.4 An omnibus approach

I introduce several changes based on the above discussion to the implementation described in Karabatsos (2001). One major difference is that I integrate the single and double cancellation checks into a single step (meaning that both instances of single cancellation as well as double cancellation are all expressed via a single jumping distribution). A form of this idea was used by Kyngdon (2011), although my double cancellation check differs substantively from his. The second major difference is how the part of the jumping distribution corresponding to the double cancellation check is formed. As I described above, I focus on just the “corners” of the 3-matrix without imposing an order through the middle cell. The final difference is that I impose the double cancellation restriction stochastically. I think this is superior to the Karabatsos approach since it treats the unknown truth value of the antecedent in a more reasonable fashion than simply assuming it to be true.

Consider a 3-matrix constructed from assessment data. Assume that the items/columns and rows/individuals have been ordered in the standard way so that the most difficult items are on the single cancellation had the necessary ordering down the minor diagonal to apply Karabatsos’s rule. So this imposition may be major in that it is occurring in all cases when only very few even warrant a test for double cancellation.
left and the least able individuals are on the top (so the probability of a correct response should increase along the major diagonal). The jumping distribution for each cell is difficult to describe due to the number of possibilities that must be considered. Remember that the goal is to establish a jumping distribution for cell \((i, j)\) that ensures both the single and double cancellation axioms are true, at least with respect to the estimates in the chain. I describe my suggested distribution first conceptually via figures and then in more precise detail following.

### 3.4.1 Conceptual Overview

Imagine we have selected a 3-matrix to examine via the checks. Karabatsos (2001) suggested we define a jumping distribution based on:

\[
\cdot \downarrow \uparrow \\
\downarrow \Theta \uparrow \\
\downarrow \uparrow \cdot
\]

where again a down arrow represents a cell whose value is less than \(\Theta\) (since it is included in the determination of the left hand bound for the uniform distribution) and an up arrow represents a cell whose value is larger than \(\Theta\). Cells with \(\cdot\) do not play a role in the jumping distribution for \(\Theta\).

Recall that my approach is meant to emphasize two facts: (1) double cancellation within a 3-matrix does not require anything in terms of the value of the middle cell and (2) double cancellation is a conditional statement, it does not need to hold in all 3-matrices.

My jumping distributions vary by location in the matrix. The middle cell is rather easy as it only has to satisfy single cancellation:\(^{16}\)

\[
\cdot \downarrow \cdot \\
\downarrow \Theta \uparrow \\
\cdot \uparrow \cdot
\]

In fact, the jumping distribution is fairly straightforward for all cells except \((1, 3)\) and \((3, 1)\) (the

\(^{16}\) In this particular 3-matrix, that is. The cell in the middle here would be the corner of many other 3-matrices.)
“corner” cells) since they only need to satisfy single cancellation. For example, for cell (2, 1) I use
\[ \downarrow \cdot \cdot \Theta \uparrow \cdot \]
\[ \uparrow \cdot \cdot \]

The complexity in the specification of the checks that I use comes in from the corner cells. While the exact details are nuanced, the fundamental idea is fairly straightforward. First consider the (3, 1) cell. This cell is involved in the consequent of the double cancellation axiom. In particular, it needs to be less in value than the (1, 3) cell if the antecedent holds. Of course, the truth of the antecedent is never known. I use a stochastic version of the antecedent for this reason. I use
\[ \cdot \cdot \uparrow? \]
\[ \downarrow \cdot \cdot \]
\[ \Theta \uparrow \cdot \]

where \(\uparrow?\) indicates that this restriction is enforced stochastically when
\[ \hat{\Theta}_{21} < \hat{\Theta}_{12} \text{ and } \hat{\Theta}_{32} < \hat{\Theta}_{23} \]

where hatted-values indicate current values in the Markov chain. Similarly for the (1, 3) cell, I use
\[ \cdot \downarrow \Theta \]
\[ \cdot \cdot \uparrow \]
\[ \downarrow? \cdot \cdot \]

where \(\downarrow?\) means that this restriction is enforced stochastically if
\[ \hat{\Theta}_{21} < \hat{\Theta}_{12} \& \hat{\Theta}_{32} < \hat{\Theta}_{23}. \]

I use an integrated single and double cancellation check for pragmatic reasons. If these two axioms can be verified (within a 3-matrix) at the same time, this would make it much more useful.

\[17\] There is actually a second version of double cancellation that needs to be checked. An updated version of the methodology that takes this into consideration is included in the Appendix. All of the results discussed in the main text use the original methodology. Analyses in the appendix suggest that there is very little difference between the results of the original and updated methodologies.
for an end-user that has measurement data and is looking to determine whether the Rasch model is a reasonable scaling method through which to create interval scales. Moreover, I have discovered that when the data are generated by the Rasch model, there are relatively few instances in which the antecedent seems to be true (based on the posterior means from a check of single cancellation). Hence, concentrating on double cancellation separately from single cancellation may be an inefficient use of time and resources. If real world examples do not seem consistent with the axioms of ACM (based on evidence provided by the checks), then interval scaling may be impossible. In that event, it may be fruitful to go back and use the single cancellation checks by themselves to determine whether ordinal interpretations are even appropriate.

3.4.2 Technical Specification

The goal is to establish a jumping distribution that is consistent with the ordering restrictions of ACM. Moreover, the jumping distribution will depend upon the current set of values in the Markov chain. Denote these current values by $\hat{\theta}_{(i,j)}$. Each jumping distribution will have the form $\text{Unif}[f(l_1, l_2, l_3), g(r_1, r_2, r_3)]$. Let us first examine the $l_k, k \in \{1, 2, 3\}$. $l_1$ and $l_2$ correspond to the single cancellation checks. Set $l_1 = \hat{\theta}_{(i,j-1)}$ unless $j = 1$ in which case we set $l_1 = 0$. Set $l_2 = \hat{\theta}_{(i-1,j)}$ unless $i = 1$ in which case set $l_2 = 0$. Finally, begin with a provisional value for $l_3$ of either $\hat{\theta}_{(3,1)}$ if $i = 1$ and $j = 3$ or 0 otherwise. This provisional value will possibly be altered via a rule described shortly. Turning to the set of values that form the right-hand end of the uniform distribution, set $r_1 = \hat{\theta}_{(i,j+1)}$ unless $j = 3$ in which case set $r_1 = 1$. Set $r_2 = \hat{\theta}_{(i+1,j)}$ unless $i = 3$ in which case set $r_2 = 1$. Finally, provisionally set $r_3 = \hat{\theta}_{(1,3)}$ if $i = 3$ and $j = 1$ or 1 otherwise.

The stochastic enforcement of double cancellation is now described. It depends on whether the antecedent in the axiom is true in terms of the current values of the Markov chain. So, if either $\hat{\theta}_{(2,1)} \geq \hat{\theta}_{(1,2)}$ or $\hat{\theta}_{(3,2)} \geq \hat{\theta}_{(2,3)}$, change the provisional estimate of $l_3$ to 0 and $r_3$ to 1. Note that due to how the provisional values are set up, this test only changes $l_3$ or $r_3$ for the (1,3) and (3,1) cells. Furthermore, for the (1,3) cell only $l_3$ is altered and for the (3,1) cell only $r_3$ is altered.
Define
\[ f(l_1, l_2, l_3) = \max\{l_1, l_2, l_3\}. \]

The definition of \( g \) is more complicated. These complexities are due to the fact that there are few guarantees about how well-behaved current values in the chain will be with respect to their ordering but it is necessary for the jumping distribution to be always well defined (by which I mean that the left hand value is strictly less than the right hand value). Begin by defining
\[
g'(r_1, r_2, r_3) = \begin{cases} 
\min\{r_1, r_2, r_3\} & : r_3 > f(l_1, l_2, l_3) \\
\min\{r_1, r_2\} & : r_3 \leq f(l_1, l_2, l_3)
\end{cases}. \tag{3.8}
\]

Note that \( g' \) will always be defined as \( \min\{r_1, r_2, r_3\} \) unless \( i = 3 \) and \( j = 1 \) (since otherwise \( r_3 = 1 \)). In that case, \( g' \) is defined in such a way to guarantee that a deviant \( r_3 \) value from the double cancellation check does not lead to a situation in which the left hand side and the right hand side of the uniform distribution “cross.” Let me illustrate using this notation:

\[
\begin{array}{ccc}
a & b & c \\
d & e & f \\
g & h & i
\end{array}
\]  

If we are forming a jumping distribution for the cell with current value \( g \), then the left hand side of the uniform jumping distribution is going to be \( f(l_1, l_2, l_3) = \max\{d, 0, 0\} = d. \) Since we are looking at the bottom left cell, we set \( r_3 = c \) and without the correction we would have \( g'(r_1, r_2, r_3) = \max\{1, h, c\} \). The punchline is that since no ordering can be assumed along the minor diagonal, it might be the case that \( c < d \). In that event, the jumping distribution could be malformed. The correction in Eqn 3.8 is meant to account for such a situation.

Now, define
\[
g(r_1, r_2, r_3) = \begin{cases} 
g'(r_1, r_2, r_3) & : g'(r_1, r_2, r_3) \geq f(l_1, l_2, l_3) \\
1 & : g'(r_1, r_2, r_3) < f(l_1, l_2, l_3)
\end{cases}. \tag{3.10}
\]

Again, if \( i \neq 1 \) and \( j \neq 3 \) then we will have \( g(r_1, r_2, r_3) = g'(r_1, r_2, r_3) \). The modification is meant to handle possible trouble involving \( l_3 \) in the case when \( i = 1 \) and \( j = 3 \). Again using
the notation in the example A.1, suppose we are forming the jumping distribution for the cell with current value \( c \). Due to the position of the cell, we set \( l_3 = g \) and the left hand side of the jumping distribution is \( f(l_1, l_2, l_3) = \max\{0, b, g\} \). Since \( r_3 = 1 \) here, the correction of Eqn 3.8 does not apply and we have \( g'(r_1, r_2, r_3) = \min\{f, 1, 1\} = f \). However, it could well be the case that \( g'(r_1, r_2, r_3) = f < g = \max\{0, b, g\} = f(l_1, l_2, l_3) \). Such an even would again lead to a malformed jumping distribution and the correction of Eqn 3.10 handles such a potentiality. Although there is some asymmetry here in that the right-hand limit is being treated slightly differently than the left-hand limit, I don’t find this especially concerning since the right-hand limits are the ones that are still to be updated at each iteration of the chain.

Once the jumping distribution is defined, the Metropolis-Hastings algorithm is used as usual. A proposed value for the parameter is drawn from the jumping distribution. If \( \hat{P}^{t-1}_{ij} \) is the last value in the MCMC chain and \( \hat{P}^*_ij \) is the proposed value, then compute

\[
r = \frac{(\hat{P}^*_ij)^{n_{ij}}(1 - \hat{P}^*_ij)^{N_{ij} - n_{ij}}}{(\hat{P}^{t-1}_{ij})^{n_{ij}}(1 - \hat{P}^{t-1}_{ij})^{N_{ij} - n_{ij}}},
\]

where \( n_{ij} \) is the number of correct responses by the individuals in row \( i \) to item \( j \) and \( N_{ij} \) is the number of individuals in row \( i \) who responded to item \( j \). Draw \( t \) from \( \text{Unif}[0,1] \). If \( r > t \), then set \( \hat{P}^t_{ij} = \hat{P}^*_ij \). Otherwise, set \( \hat{P}^t_{ij} = \hat{P}^{t-1}_{ij} \). In each cell of the 3-matrix, I perform 3000 iterations of which the first 1000 are burned and then the chain is thinned to only every fourth draw. I then estimate the 95\% credible region for the parameter using the remaining values. A distortion is detected here if

\[
\hat{P}_{ij}^{\text{MLE}} = \frac{n_{ij}}{N_{ij}}
\]

does not lie in this 95\% credible region. Remember that 9 Markov chains are being formed simultaneously within each 3-matrix.

I wrote code to perform the checks as described here for \( R \) (R Development Core Team, 2010). Since each choice of 3-matrix requires a set Metropolis-Hastings runs, choosing thousands of 3-matrices as described in subsequent chapters can be quite computationally intensive. However, this type of algorithm is “embarrassingly parallel” since it can done in parallel with no additional
work (each 3-matrix can be handled separately in an independent process). I used the snow package (Tierney, Rossini, Li, & Sevcikova, n.d.) to parallel-ize the analysis. Finally, I used the Research Computing platform at the University of Colorado at Boulder and leveraged large numbers of processors into reasonable run times. As a performance benchmark, I was able to use around 100 processors to check over 10,000 3-matrices (that’s 90,000 Markov chains) in under an hour using this system.

3.5 The Fit of the Rasch Model

If the Rasch model is the true data generating process, then the axioms underlying ACM will hold and interval scales can be created. This is not too difficult to demonstrate. Consider persons with abilities $\theta_1, \theta_2, \text{ and } \theta_3$ to items with difficulties $b_1, b_2 \text{ and } b_3$ where $\theta_1 \leq \theta_2 \leq \theta_3$ and $b_2 \leq b_2 \leq b_3$. We can explicitly state the true probability of a correct response. For $i, j \in \{1, 2, 3\}$ we have

$$P_{ij} = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}.$$

Single cancellation is equivalent to the Rasch concept of specific objectivity. Double cancellation is obtained as follows. First, we apply the following monotonic transformation

$$\log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \theta_j - b_i.$$

We now note that if both $\theta_2 - b_1 \leq \theta_1 - b_2$ and $\theta_3 - b_2 \leq \theta_2 - b_3$ are true then we can use the same steps to show $\theta_3 - b_1 \leq \theta_3 - b_1$ as before (and where we define $\phi_1(x) = x$ and $\phi_2(y) = -y$). Since abilities and item difficulties can take any value along the real number line, solvability holds as a consequence of the density of the reals and the Archimedean condition holds since a similar condition holds for the real numbers.

---

18 This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research.

19 Parts of this argument are derived from Borsboom (2005).

20 This argument is perhaps difficult to see at first due to the finiteness of any given measurement instrument. Since there are only a finite number of items (for example), how can the existence required by solvability be guaranteed? What is required, however, is not that such an item does in fact exist but that one could. Consider momentum: the fact that one can have only a finite number of balls moving at a finite number of speeds does not raise questions about whether an interval scale is possible. In the case of the Lexile scale, writing an item of a specific difficulty is
Some have argued that fit tests are not necessarily good indicators of agreement with the axioms of ACM. Perline et al. (1979) examined this question and Karabatsos (2001) also contains some thoughts on the subject. In general any statistic computed on the basis of estimated parameters or residuals should be viewed suspiciously as the parameters are estimated and yet treated as true and the residuals are minimal with respect to some criteria. Research has shown that even in the case of axiomatic violations certain tests conclude perfect model fit (Nickerson & McClelland, 1984). They demonstrated that fit statistics based on numerical conjoint measurement techniques were unable to detect known distortions.²¹ Even statistics not based on estimated parameters (such as the scalability coefficient I discuss below) may suffer from such problems, although such problems are harder to identify. For all these reasons, the Rasch model should not be the initial indicator as to whether interval scales are possible with a given data. First, the checks will be used to determine whether the axioms of ACM are consistent with the data. In a second, separate step, I will then analyze the Rasch model fit. If the checks seem to support the axioms, then the question will be whether the Rasch model fits. If so, the resultant scale should be interval.

There is a substantial literature around the concept of “model fit” as it relates to the Rasch model; I will not review the entire scope of this research body but merely touch upon the relevant aspects. Traditionally it has been possible to divide fit statistics into two types: those based on individual-level residuals (differences between actual responses and estimated probabilities of a correct response) and those based on the responses of groups aggregated on some factor, typically common sum scores (this last type includes the many likelihood ratio tests in Glas & Verhelst, 1995²²). In this study, I will concentrate on several statistics some of which do not fit so easily into a relatively easy endeavor as one just need to write a cloze item of requisite length and complexity. Interestingly enough, momentum does demonstrate an example of how solvability can fail but this is due to other considerations. Since velocity is bounded from above (due to the theory of relativity), it seems plausible to imagine an incredibly massive object moving at a slow speed with a momentum that can not be matched by an object with far less mass due to the limit on velocity.

²¹ Although Nickerson and McClelland (1984) was not concerned explicitly with Rasch fit statistics, it is worth noting their conclusion that fit statistics have been shown to be flawed from the perspective of ACM.

²² The likelihood ratio tests discussed in Glas and Verhelst (1995) are discussed mainly as tests for DIF since they are based on parameters estimated separately over several subgroups. For this reason, I do not see them as being especially useful for my purposes (note also that they do not fair as well as $H^1$ in Karabatsos (2001)). While it is the case that almost any test statistic for the Rasch model could potentially be used for this project, I have chosen to focus on those that are sensitive to violations of the Rasch model that are similar in spirit to violations of the
this residual/group dichotomy since they are specifically concerned with anti-Guttman responses.

The first statistics are the infit mean squares and outfit mean squares statistics based on the work of B. Wright and Panchapakesan (1969).\footnote{These are the standard fit statistic for Rasch scales and thus merit inclusion even if they are not specifically related to axiom violations.} I shall discuss all statistics as “person-fit” statistics but by a changing of indices they can easily become “item-fit” indices. Each mean square statistic begins with a rescaled residual

\[ Z_{ij} = \frac{Y_{ij} - P_{ij}}{\sqrt{P_{ij}(1 - P_{ij})}} \]

where I use the same notations as in Eqn 2.1. Note that the denominator above is the variance of a Bernoulli random variable. The outfit version (OMS) is then formed by taking the sample average of the squares for an individual over \( I \) items

\[ OMS_j = \frac{\sum_{i=1}^{I} Z_{ij}^2}{I}. \]

The infit version (IMS) is computed using a weighted average

\[ IMS_j = \frac{\sum_{i=1}^{I} Z_{ij}^2}{\sum_{i=1}^{I} P_{ij}(1 - P_{ij})}. \]

Wu and Adams (n.d.) showed that these statistics serve as tests of the assumption about whether there does in fact seem to be a single unique slope for all of the item characteristic curves. The critical values in Wu and Adams are also useful since they are computed as a function of sample size.

I shall also consider a form of the statistic suggested by Van der Flier (1977, as cited in Karabatsos, 2000). In its most basic version this statistic has the form

\[ G^*_j = \frac{\sum_{k<i} Y_{ij}(1 - Y_{kj})}{r_j(I - r_j)} \]

where \( r_j = \sum_{i=1}^{I} Y_{ij} \) is the person sum score. In the numerator, item \( i \) is the more difficult of the item pairs under consideration. Note that if the items in a given pair are both answered correctly axioms of conjoint measurement. However, it is not difficult to imagine that statistics designed to test other violations (such as DIF or unidimensionality) could be interesting since such violations may lead to non-parallel item response functions. This project is not meant to be an exhaustive survey of Rasch model fit indices, so I am working to select only those indices most likely to be useful.
or incorrectly then nothing is contributed to the statistic by this pair. If the easier item is answered correctly and the harder one incorrectly, again nothing is contributed. It is only when the easier item is answered incorrectly, $Y_{kj} = 0$, but the harder item is answered correctly, $Y_{ij} = 1$, that anything is contributed to the numerator. This statistic is simply a weighted count of the number of item pairs that are anti-Guttman. It is especially appealing here as it “is essentially a conjoint additivity test for a vector of (person or item) responses” (Karabatsos, 2000, p. 171). More recent versions (Rost & von Davier, 1994) take into account the magnitude of the anti-Guttman like response by weighting with the item difficulty $b_i$:

$$Q_j = \frac{\sum_{i=1}^{I}(Y_{ij} - Y_{iG})b_i}{\sum_{i=1}^{I}(Y_{iA} - Y_{iG})b_i}.$$  

$Y_{iG}, Y_{iA} \in \{0, 1\}$ are the expected Guttman and anti-Guttman responses where these are identified as the maximum and minimum likelihood responses for person $j$ to items $i \in \{1,...,I\}$ when conditioned on sum scores. $Q_j \in [0,1]$ provides information about item discrimination: a value of 0 indicates perfect discrimination while 1 indicates perfect negative discrimination (paraphrased from Rost & von Davier, 1994, p. 176 where guidelines for statistical inference are also given).

Finally, the $H^T$ statistic of Sijtsma and Meijer (1992) shall be used due to its success in the simulation study discussed in Karabatsos (2001). This statistic is based on Guttman-type errors if the “more able person $j$ answers an item, say $g$, incorrectly, and person $i$ ($\beta_i \leq \beta_j$) answers item $g$ correctly” (Sijtsma & Meijer, 1992, p. 151). For individual $i$, the statistic is defined as

$$H^T_i = \frac{\sum_{j \neq i} \sigma_{ij}}{\sum_{j \neq i} \pi_j (1 - \pi_i)}$$

where $\pi_i$ is the proportion of items that individual $i$ responded correctly to, $\sigma_{ij}$ is the covariance between scores on items by individuals $i$ and $j$ (that is, $\sigma_{ij} = \pi_{ij} - \pi_i \pi_j$ where $\pi_{ij}$ is the proportion of correct responses by both $i$ and $j$), and $j$ indexes all individuals besides $i$. Note that the denominator leads to the maximal covariance between the two individuals. Whereas all of the other fit statistics were computed using code written by the author, van der Ark (2011) was used to compute this “scalability” coefficient.\(^{24}\)  

\(^{24}\) I used other packages of IRT software (Rizopoulos, 2006; Mair, Hatzinger, & Maier, 2011; Partchev, 2011).
computed in this way can be used to determine whether a set of items can be scaled together.
Statistical tools are not necessarily endowed with their advertised features. The choice of a 95% credible region from the posterior density suggests that we may expect to detect distortions in roughly 5% of the checks where the underlying data are consistent with the Rasch model. But this expectation should be checked. While item-side deviations should lead to distortions, not all distortions may be due to item-side deviations. Other possible causes of distortions are random error or person-side deviations.\footnote{One important piece of terminology: I use “distortion” to mean either problems detected in the response data (via either the fit statistics or the checks) or problems with the scale created from the response data. In contrast, I use “deviation” to indicate item parameters deviating from the Rasch model.} Furthermore, it is natural to expect the capability of the checks to detect distortions induced by deviations from the Rasch model to depend upon the size of the deviation.

This chapter is guided by several specific research questions. First, are the statistical properties of the checks such that they can be reliably expected to detect data suffering from such item-side deviations ([Q1] from Chapter 1)? This question was not answered in either Karabatsos (2001) or Kyngdon (2011) and the entire methodology will be of limited utility if the answer is in the negative. Second, how does the information provided by the checks differ from the information provided by the fit statistics [Q1A]? Since the latter are already well known tools in psychometrics, if the checks add nothing to our understanding then it would be unclear they are worth the additional trouble. Third, what is the impact of known item-side deviations on the estimated scale [Q2]? In particular, what effect would the distortions have on how we use educational assessment
data in the real world.

I begin with “size” and “power” analyses that give baseline information about the performance of the checks in simple situations. Note that these analyses differ slightly from what is traditionally meant by size and power analyses in that I am simultaneously checking hypotheses regarding a number of parameters rather than focusing on a single parameter. Once the behavior of the checks has been analyzed in these simple settings, I move into a pair of simulations with more complex sets of distortions. This pair of simulations is again based on item-side deviations, but they differ in important respects from one another in terms of how the data were simulated. Finally, I examine the potential impact of mistakenly assuming an interval scale (created via the Rasch model) exists (when the data was in fact generated by the 3pl). Although this question is quite important, it turns out to be quite challenging to answer.

4.1 Basic Simulation

4.1.1 “Size” Analysis

The discussion of the statistical properties of the checks begins with a look at their performance when the data is simulated via the Rasch model. In such a case, there are no violations of the measurement axioms and distortions should be detected at low levels. The precise level is connected to the choice of credible region for the posterior from the previous chapter. Recall that I am working with the 95% credible region of the posterior. Data was simulated via the Rasch model for 15,000 respondents and 45 items. These were chosen to be reflective of the size of data one may see working with a smaller state. The abilities are random draws from the standard normal. The item difficulties are evenly distributed between -1.5 and 1.5 (-1.5, -1.4318, -1.3636, ...; note that the difference between any two item difficulties is constant). This was done to reflect a certain ideal case for Rasch measurement. In such a case, the measurement instrument is as much like a typical ruler as can be hoped for. I then choose 8,000 3-matrices to check for distortions.\footnote{Choosing an appropriate number of 3-matrices is tricky. Without enough 3-matrices, a reasonable overall assessment of violations may not be established. With too many, computational times soar. I address the choice for the number of 3-matrices by comparing the rate of detected violations after increasingly larger choices for the number of 3-matrices.} The units of
analysis are the cells corresponding to a single item and a set of individuals at a unique sum score.

When the hypothesized model is true, the method under-detects distortions relative to this 5% threshold and the detection rate is influenced by the position of the cell in the conjoint matrix. The overall detection rate is rather low, only 1.3%. It is tempting to immediately compare this to the standard 5% threshold, but a little caution must be taken in doing so. In standard usage, this threshold is related to a single test. By this, I mean that only a single parameter is being tested when we normally consider alpha-levels, confidence intervals, and Type I errors. Here, focus in on the percentages of violations detected in a multitude of tests (9 tests in each 3-matrix, of which I’m sampling 8,000!) where detection is based on a 95% credible region.

Intuitively, a smaller credible region (e.g., 90%) should lead to more detected violations since fewer of the observed proportions of correct responses will fall within these regions (simply since they are narrower). To demonstrate this, I simulated Rasch data consisting of 20 items (difficulties evenly spaced between -1.5 and 1.5) and 2000 respondents and computed the number of detected violations based on 4,800 sampled 3-matrices for 90%, 95%, and 99% credible regions. The percentages of detected violations were 3.6%, 2.1%, and 1.0% respectively. That this corresponds with the stated intuition is encouraging and suggests that more restrictive or lax tests of the cancellation axioms could be performed by appropriate choice of credible region. Here, it is prudent to consider the 1.3%, based on a 95% credible region, as a baseline that will prove useful in later discussions when we want to compare non-simulated data back to Rasch data.

Figure 4.1 demonstrates that there is a great deal of variability in the proportion of detected violations at given item/ability combinations. The figure shows the proportion of violations on the y-axis. The x-axis shows sum scores and each line represents a unique item. This means that over any sum score, there should be 45 distinct points representing the 45 items. The fact that there appear to be far fewer is due to the number that are lying on the x-axis. There is a dashed horizontal line at the 5% threshold and clearly many items break this threshold at some point but

---

A discussion of this issue is postponed until later in this Chapter since a determination of the number of 3-matrices based on the true Rasch simulation may be too low. In this simple setting, 8,000 should be a sufficient number.
Figure 4.1: Proportion of violations for items at each sum score when underlying model is truly Rasch
there is a reason to overlook these violations: they are largely localized. There are few (if any) instances in which violations are detected at a high rate for several sum scores in a row (within a single item) which we would expect if an item was truly behaving aberrantly in a specific ability region. It is important to note that there are more detected violations at the extreme abilities. This is even after I have only worked with sum scores with more than 25 (out of the 15,000) respondees. This initial exploration demonstrates that the checks behave reasonably when the data are truly Rasch. The obvious next question would be about the performance of the checks when the Rasch model is not the true model.

4.1.2 “Power” Analysis

In this section I present the results from two simple analyses while also demonstrating the patterns of distortions that the checks follow under certain known deviations from the Rasch model (based on either the occurrence of guessing or items with differential discriminations). The two simulations in this section are based on 15,000 abilities drawn from the standard normal and 45 items with difficulties evenly distributed from -1.5 to 1.5. In each simulation, a single item’s parameters are made to deviate from the Rasch model (by adding either a guessing or non-unity discrimination parameter) and the resulting data is then checked for distortions.

4.1.2.1 Guessing

In Figure 4.2, I examine the performance of the checks in identifying an item where guessing is known to occur, but the other 44 items are purely Rasch. In each figure, the item with guessing is highlighted in red and the true difficulty and guessing parameter for that item are shown at the bottom. Beginning on the left hand side, the guessing parameter goes from 0.1 to 0.4 in increments of 0.1. Going down the first 3 panels on the left hand side, one can easily detect a great increase in the proportion of violations detected for the item in red. However, there is an important interaction occurring that masks this effect in the bottom left panel. While we would expect to see more distortions associated with larger guessing parameters, there is a strong dependency on
the item difficulty. In the middle two panels on the left, the distortions are easy to detect in part because of the fact that these are difficult items. In the bottom left panel, the problems seem minor due to the ease of the item. On the right hand side, I’m using the same guessing parameter (0.25) but choosing easy, medium, and hard items. Note that the medium and easy item (top two) show basically no distortion. On the difficult item, more students are in the region of the ability spectrum where guessing becomes a stronger possibility and this is resulting in the easily identified distortion.

## 4.1.2.2 Discrimination

Figure 4.3 is similar to Figure 4.2, but here the discrimination is varying and no items exhibit guessing. In the left column, the discriminations are fixed at 0.5, 0.75, 1.25, and 1.5. Distortions are not visible for the 0.75 and 1.25 discriminations, but the more extreme discriminations are easy to identify due to the bulges on the left and right hand side of the red curve whereas the red curve is near 0 in the middle. This is due to the fact that the cells in the middle (in terms of sum scores) did not have the probability of a correct answer altered dramatically by the distortion. In terms of the right column, where the same discrimination of 0.75 is paired with multiple different item difficulties, there is not much of a discernible pattern when the discrimination is fixed at 0.75 and the difficulty is varied but this is mostly due to the fact that 0.75 is not a sufficient deviation to induce major distortions.

There are two key points from these first 3 simulations to recall going forward. First, the checks seem to under-detect distortions relative to a 5% threshold in the case of truly Rasch data. This will be a useful fact to recall in empirical examples. Second, the checks do seem to detect simple distortions. The deviations I’ve used here—varying the guessing and discrimination parameters—are certainly not the only possible deviations from the Rasch model, but are likely candidates for problems when multiple choice data is scaled using the Rasch model. This initial exploration was relatively simple, but an important demonstration of the functionality of the checks. The following section describes a more complex simulation where a variety of deviations are induced
Figure 4.2: Proportion of violations for items at each sum score when guessing is varying.
simultaneously.

4.2 Reflections on first simulations

In the first set of simulations, we’ve seen that the checks were able to identify single items that deviated from the Rasch model. It seems unlikely that real world data to be scaled using the Rasch model would have only a single item deviating from the Rasch model (assuming that the assessment has not already been pruned of deviant items). Rather, most items will probably deviate from the Rasch model in some respect and these deviations will be of varying degrees. It is not yet known how these deviations will be reflected by the findings from the checks. Understanding the relationship between the checks and more complex sets of deviations is one of the goals of the next section.

There are two factors in particular that seem important in setting up subsequent simulations. The first is how to simulate abilities. It is not immediately apparent how to best simulate abilities in this context. The intuitive approach would be to sample a large number of abilities independently from the standard normal. Let’s say 15,000 abilities were sampled in such a manner. This yields 15,000 distinct abilities. Applying the Rasch model to the subsequent assessment data yields a maximum of \( n + 1 \) distinct ability estimates on a test with \( n \) items. It is clearly unreasonable to expect the Rasch model to reliably replicate all of the scale properties for the 15,000 true abilities when it is capable of generating only a much coarser set of estimates.\(^3\) Moreover, it seems unlikely that we’d even be interested in being able to measure abilities at that level of precision in an educational setting. How abilities are simulated is the first factor that I vary in the two versions of the second simulation.

There is a second decision to be made with respect to the simulation to follow. One natural question is how well the estimated abilities resemble the true abilities. In the case of data simulated according to the Rasch model, we don’t expect recovery to be perfect due to measurement error. If

\(^3\) This assumes that \( n \ll 15000 \), a reasonable assumption as tests with several thousand items would be even more monotonous than reading this dissertation.
Figure 4.3: Proportion of violations for items at each sum score when discrimination is varying
data is simulated using a more complex model such as the 3pl, then there are going to be two types of “error”. Measurement error as conventionally understood as well as additional scale distortions due to the violations of the measurement axioms. One must account for both in comparing true and estimated abilities. This is the second factor that varies in the following simulations.

I take the following approach towards these two factors. In the first version of the second simulation, abilities are simulated in what I described as the intuitive way and estimated and true abilities are compared without explicitly trying to account for the two types of error just described. In this first version of the simulation, focus is on the performance of the checks as well as several other questions (e.g., how do fit statistics compare to the checks). In a redux of this simulation, I simulate abilities in a manner that leads to a more reasonable number of ability levels. In this redux, I also compare abilities estimated with non-Rasch data to the ability estimates from Rasch data (although the underlying true abilities are the same in each case, the items differ in that additional parameters are introduced in the non-Rasch case). Since the Rasch data conforms to the measurement axioms, the scale should be interval and the resulting estimates are a best-case scenario for producing an interval scale. The focus in the redux will not be on the performance of the checks, but on the impact on the ability scale that results from these modified assumptions.

4.3 Second Simulation

This simulation is meant to analyze the performance of the checks in a more realistic setting than the basic simulations discussed earlier in this chapter. Alongside the checks, I shall also consider item fit statistics as a gauge of the suitability of the Rasch model for a given set of data. There are a variety of conditions that I’ll vary in this simulation, but it is always based upon 15,000 students (abilities drawn from $N[0, 1]$) responding to 45 dichotomously scored items (with difficulties in a evenly spaced sequence between -1.5 and 1.5). The key variant in this simulation is the quantity of items (0, 10, 20, or 40) in which there is either guessing or a non-unity discrimination. The full crossing of these two sets of item-side deviations from the Rasch model results in 16 simulation conditions. When the guessing parameter is not fixed at 0, it is sampled from a uniform distribution
When the discrimination is not fixed at unity, it is sampled from a uniform distribution on \([0, 0.3]\). Compared to the deviations from the Rasch model observed in empirical data from Colorado (detailed below), these deviations are about what one would expect to see in practice with data traditionally thought to necessitate the more complex 3pl model. Interestingly, these are fairly modest deviations compared to the variation in discrimination and guessing observed in the empirical example shown in Figure 5.4, data that is scaled with the Rasch model. Before discussing the results of the simulation, let me first explain the rationale behind the number of 3-matrices (16,800) which I sampled from the data matrix to check for distortions.

### 4.3.1 Number of sampled 3-matrices

The choice of the number of 3-matrices is a trade-off between the computational time required and desired accuracy. Remember that to perform the checks on a single 3-matrix corresponds to creating 9 different Markov chains via the Metropolis-Hasting algorithm. No small task in and of itself and to do this for a large number of 3-matrices could clearly take a while. For the second simulation, I chose 16,800 3-matrices per iteration of the simulation. The reason for this particular number is partially a function of the structure of the Janus supercomputer which has 12 cores per node. As a rough estimate, suppose that the 45 items lead to 40 sum scores with reasonable numbers of individuals (25 out of the 15,000). This is a total of 1,800 cells in the resulting conjoint matrix. Each 3-matrix contains 9 cells so there should are \(9 \cdot 16800\) cells drawn in total. Hence, each cell will be drawn for inclusion in a 3-matrix

\[
\frac{9 \cdot 16800}{1800} = 84
\]

84 times on average.

Is this a sufficient number of draws to get a stable estimate of the percentage of violations that one would detect in a cell if one could check every 3-matrix (a test of the size of the one simulated here would have somewhere in the neighborhood of \(10^8\) 3-matrices)? To answer this, I re-ran the iteration of the simulation with 20 items distorted by guessing and 20 items distorted...
by differential discriminations and drew 100,000 3-matrices. I then computed the percentage of 
cells showing distortions for 5,000 3-matrices, 10,000 3-matrices, and so on all the way to the full 
draw of 100,000. Figure 4.4 shows these percentages (for each sum score cell within an item) over 
the increasing number of drawn 3-matrices. Ideally what we’d see is the lines becoming flatter 
(meaning that they are stably estimating the proportion of violations at that ability level), which 
they do. The red vertical line represents the number of draws used in the second simulation. For 
this to be an insufficient number of matrices, one would need to see substantial changes occurring 
after this line over a large number of items. This is not the case so something around 15,000 draws 
is a reasonable number of draws for a conjoint matrix of this size.

It is also important to remember the difference between what I am doing and what was done 
in, for example, Kyngdon (2011). In that article, he only checked two overlapping 3-matrices. This 
means that only 12 cells of the (40 ability levels times 54 items) 2160 cells in the full data matrix 
were examined at all. Not only would my approach involve each of the 2160 cells (I perform such 
an analysis on the same data, with very different results, in the next chapter), but each cell would 
be tested in a variety of 3-matrices, some of which may show violations while others might not since 
a violation depends not on a single cell alone, but upon the choice of other cells with which it is 
compared. Even if the chosen number of sampled 3-matrices (16,800) seemed insufficient to stably 
estimate what one would find if all 3-matrices could be checked, the repeated sampling approach 
would still be a serious methodological improvement over checking only some paltry number of 
3-matrices.

4.3.2 Results

I begin the discussion of the results using Table 4.1 which documents the overall performance 
of the checks across all 16 conditions of the simulation. The two columns “N Disc” and “N Guess” 
show the number of items (out of the 45) that had either a non-zero lower asymptote to represent 
guessing or a discrimination not equal to unity. The remaining columns give various means for the

---

4 Curves are piecewise linear between each multiple of 5000.
Figure 4.4: Stability of the proportion of checks showing distortion over varying numbers of draws. Vertical red line represents 15,000 draws while the x-axis extends all the way to 100,000 draws. Y-axis shows percentage of violations detected for each item. Note that, in general, after the lines cross the vertical red line they have flat slopes.
percentage of detected violations. The “Mean” columns shows the mean detection rate across all cells. The “Wt Mean” column is a weighted mean where each cell is weighted by the number of respondents with that particular sum score. (Note that as in the previous sections, all work is done with sum scores containing at least 25 individuals.) This is meant to downplay the volatility that we see at the margins, caused due to the additional error in estimating probabilities at the extreme abilities since there are so few respondents. The last 3 columns show weighted means for different types of items. The “Wt Mean-Disc” column shows the weighted mean for only those items with non-unity discriminations. The next two columns display weighted means for those items that have a non-zero lower asymptote (“Wt Mean-Guess”) and for those items that are truly Rasch (“Wt Mean-OK”).

The first row demonstrates the performance of the checks in a truly Rasch setting. Here, across all cells, we’re seeing roughly 2% of the checks concluding that there are distortions. Under the weighting, this falls to 1%. These are in the ballpark of what was observed in the size analysis above. In the first four rows, we’re seeing increasing number of items affected by guessing but with all equal discrimination. Interestingly, when 40 items have guessing, the items without guessing actually get flagged more frequently for distortions than the items with guessing. If we compare those rows to rows 1, 5, 9, and 13 (where there are increasing numbers of items with non-unity discriminations but no guessing), the checks are better able to detect deviant discriminations than guessing. Moreover, the additional detection of discrimination does not come with a steep uptick in the number of “OK” items showing detections. Looking at row 16, we can see that in a situation with substantial amounts of guessing and differences in discrimination, all types of items are consistently being flagged relative to the 1% baseline (determined in the purely Rasch case).

Focusing on the changes from row to row in Table 4.1, one might view the changes in the weighted mean as only slight increases of a few percentage points. But that ignores the fact that the increases are dramatic relative to one another. Relative to the weighted mean of the truly Rasch condition, we see increases of 305%, 511%, and 680% for the increasing item-side deviations (again, looking at the simulations with the same number of guessing and discrimination deviations). Or, if
Table 4.1: Summary table for second simulation: Mean % of detected distortions as a function of the simulation conditions.

<table>
<thead>
<tr>
<th>N Disc</th>
<th>N Guess</th>
<th>Mean</th>
<th>Wt Mean</th>
<th>Wt Mean-Disc</th>
<th>Wt Mean-Guess</th>
<th>Wt Mean-OK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>02</td>
<td>01</td>
<td>02</td>
<td>01</td>
<td>01</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>10</td>
<td>02</td>
<td>01</td>
<td>02</td>
<td>01</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>20</td>
<td>03</td>
<td>02</td>
<td>02</td>
<td>02</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>40</td>
<td>04</td>
<td>02</td>
<td>02</td>
<td>03</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0</td>
<td>04</td>
<td>02</td>
<td>06</td>
<td>01</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>10</td>
<td>04</td>
<td>03</td>
<td>05</td>
<td>03</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>20</td>
<td>05</td>
<td>03</td>
<td>06</td>
<td>03</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>40</td>
<td>05</td>
<td>03</td>
<td>06</td>
<td>03</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>0</td>
<td>05</td>
<td>04</td>
<td>05</td>
<td>02</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>10</td>
<td>05</td>
<td>04</td>
<td>05</td>
<td>04</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>20</td>
<td>07</td>
<td>04</td>
<td>06</td>
<td>04</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>40</td>
<td>06</td>
<td>04</td>
<td>06</td>
<td>04</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
<td>0</td>
<td>08</td>
<td>06</td>
<td>06</td>
<td>02</td>
</tr>
<tr>
<td>14</td>
<td>40</td>
<td>10</td>
<td>08</td>
<td>06</td>
<td>06</td>
<td>05</td>
</tr>
<tr>
<td>15</td>
<td>40</td>
<td>20</td>
<td>09</td>
<td>06</td>
<td>07</td>
<td>06</td>
</tr>
<tr>
<td>16</td>
<td>40</td>
<td>40</td>
<td>08</td>
<td>06</td>
<td>06</td>
<td>05</td>
</tr>
</tbody>
</table>
one prefers, relative to the case with 10 items that have guessing and 10 items that have non-unity
discriminations, we see increases in the detected distortions of 168% and 223%. Compared to these
baselines, these are noteworthy differences that are easily detected visually in Figure 4.5.

Despite these differences, it may seem concerning that even when a great number of deviations
from the Rasch model exist it is still only a small proportion of the checks that detect violations.
However, this concern would be misguided. The checks are primarily concerned with local, not
global, behavior. As a consequence, the item deviations that I use induce distortions in certain
locales which then must be selected for examination through choice of a specific 3-matrix (both the
item with deviation and the other items selected for the 3-matrix matter in detecting a distortion).
All of which is to say that distortions, even for items with deviations, are not going to be detected in
all 3-matrices with the problematic item. Based on the empirical evidence, it seems that detecting
such violations is relatively rare. Another way to consider these changes is by the looking at the
number of ability groupings for each item which show any violations. For the truly Rasch case,
violations were detected (in any case) for 11.3 of the ability groupings for the average item. Looking
again at those simulation conditions with equal numbers of item-side deviations, these means rise
from 11.3 to 18.8, 24.5, and 26.3 respectively for 10, 20, and 40 distortions. The checks clearly
detect violations, but it requires potentially looking at large numbers of 3-matrices.\(^5\)

Figure 4.5 shows the performance of the checks in selected iterates of the simulation where
there are equal numbers of items with non-zero lower asymptotes and non-unity discriminations. At
the bottom of each figure, “D” represents the number of items with differential discriminations and
“G” represents the number of items with guessing. This set of iterations (same number of items with
guessing and difficulty deviations) will be used frequently in the below analyses.\(^6\) As the number
of items with item-side deviations increases, we are clearly seeing increased detection of distortions
from the checks. At first glance, it may seem that the information here differs dramatically from
Table 4.1 (e.g., way more distortions seem to be shown in the Figure than are reflected in the means

\(^5\) In other words, looking at only a few 3-matrices as in Karabatsos (2001) and Kyngdon (2011) is unlikely to be
effective.

\(^6\) Going clockwise from top left in Figure 4.5, these represent rows 1, 6, 16, and 11 respectively from Table 4.1.
Figure 4.5: Performance of checks in simulation iterations with same number of items that have guessing and differential discrimination. Black lines indicate truly Rasch items while red lines indicate items that have either guessing or non-unity discriminations.
of the Table). That is not the case, but there are some facts that one must remember to reconcile the two. Note first that the nature of the distortions that I introduce on the item side are such that they should only induce distortions detectable by the checks in certain places. For example, when I introduce guessing into an item, this should not alter probabilities for high-ability students in a meaningful way. In contrast, discriminations that are quite different from unity still do not alter the responses by individuals whose true ability is close to the true item difficulty. So even for deviant items, distortions should not be detected in all cells.

Furthermore, in all of the iterates shown in Figure 4.5, a large number of the cells show no distortions. As a consequence, there are a large number of over-plotted lines on the x-axis. Going from least to most item-side deviations, 71%, 50%, 36%, and 26% of the cells in each instance have no detected distortions (and remember that each cell is getting sampled upwards of 80 times on average). These weight the means in Table 4.1 heavily but are basically undetectable in Figure 4.5. For these reasons, these figures must be interpreted with some care. Despite this, there is still tremendous utility in these figures when they are used for comparisons. For example, the increase in item-side deviations is apparent when one goes from the top left of Figure 4.5 to the bottom right. Similarly, they illustrate the problematic items nicely in Figures 4.2 and 4.3. For these reasons, I shall continue to use them even if they come with some caveats.

4.3.3 Comparing fit statistics to the checks

The checks are meant to indicate whether the data is consistent with the measurement axioms laid out in prior chapters. Recall that item fit statistics are meant to determine whether response data is consistent with the Rasch model. Since any data that is consistent with the Rasch model will necessarily obey the measurement axioms, it is natural to inquire as to whether the fit statistics and the checks generate the same kind of information. To compare the fit statistics and the checks, I focus on only those iterates of the simulation with the same number of items that have guessing and differential discriminations.
Figure 4.6 shows a selection of items from each of these four iteration. The left most column shows items from the iteration of the simulation in which there were no deviations from the Rasch model and the right most column the simulation with 40 items of each type of deviation. Interpreting each individual panel is a two-step process. The first step pertains to the red line. The red lines in each figure show the percentage of checks (using the cell created by a given ability level and the relevant item) that detected distortions. This Figure is similar to what was first shown in Figure 4.1 or 4.5 but with just one item per panel. These can all be judged relative to the faint dotted red line which is at 0.05, but it is more informative, in my opinion, to focus on changes across the ability level and changes within an item across the simulation.

The second step in interpreting each panel pertains to the four characters. The characters (o, i, q, and s) represent the standardized values of four fit statistics: outfit mean squares, infit mean squares, item q-index (Rost & von Davier, 1994), and the scalability coefficient (Sijtsma & Meijer, 1992). These can be compared to the higher horizontal line which is a baseline of sorts for the fit statistics. In each iterate of the simulation, the fit statistics are standardized relative to the mean and SD of the first, truly Rasch, iterate of the simulation. The scale is such that small deviations from this line show relatively large deviations from zero (that is, if the character doesn’t overlap with the baseline than it is showing relatively serious misfit). To understand the scale, note that the standardized outfit fit statistics for item 2 are at -.1, -.2, -.4, and 1.1. The final piece of information in each panel are the item parameters which are at upper right. These are discrimination, difficulty, and guessing from top to bottom (all item parameters for the iteration with 40 deviant items of each type are shown in Table 4.2).

The left column is from the condition with no item-side deviations from the Rasch model. As a consequence, we see quite little in the way of distortions in either the statistics or the checks. To understand the difference in the information delivered by the checks and the fit statistics, consider item 14. In the second column, this item begins to have responses affected by guessing. This causes the noticeable bump on the left hand side of the line corresponding to the checks. The fit
Figure 4.6: Small Multiple Figure comparing checks to fit statistics
Table 4.2: Item Parameters from iteration of simulation with 40 items with non-zero guessing parameter and 40 items with non-unity discrimination

<table>
<thead>
<tr>
<th>Discrimination</th>
<th>Difficulty</th>
<th>Guessing</th>
<th>Discrimination</th>
<th>Difficult</th>
<th>Guessing</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.79</td>
<td>-1.50</td>
<td>0.26</td>
<td>1.40</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>1.00</td>
<td>-1.43</td>
<td>0.04</td>
<td>1.20</td>
<td>0.14</td>
<td>0.00</td>
</tr>
<tr>
<td>0.57</td>
<td>-1.36</td>
<td>0.29</td>
<td>0.64</td>
<td>0.20</td>
<td>0.27</td>
</tr>
<tr>
<td>1.00</td>
<td>-1.30</td>
<td>0.00</td>
<td>1.32</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>0.52</td>
<td>-1.23</td>
<td>0.03</td>
<td>0.80</td>
<td>0.34</td>
<td>0.18</td>
</tr>
<tr>
<td>0.66</td>
<td>-1.16</td>
<td>0.23</td>
<td>1.35</td>
<td>0.41</td>
<td>0.09</td>
</tr>
<tr>
<td>0.90</td>
<td>-1.09</td>
<td>0.20</td>
<td>1.35</td>
<td>0.48</td>
<td>0.14</td>
</tr>
<tr>
<td>1.16</td>
<td>-1.02</td>
<td>0.30</td>
<td>0.61</td>
<td>0.55</td>
<td>0.12</td>
</tr>
<tr>
<td>1.48</td>
<td>-0.95</td>
<td>0.19</td>
<td>1.48</td>
<td>0.61</td>
<td>0.29</td>
</tr>
<tr>
<td>0.95</td>
<td>-0.89</td>
<td>0.08</td>
<td>0.99</td>
<td>0.68</td>
<td>0.00</td>
</tr>
<tr>
<td>1.14</td>
<td>-0.82</td>
<td>0.12</td>
<td>1.38</td>
<td>0.75</td>
<td>0.17</td>
</tr>
<tr>
<td>1.24</td>
<td>-0.75</td>
<td>0.21</td>
<td>1.06</td>
<td>0.82</td>
<td>0.17</td>
</tr>
<tr>
<td>0.69</td>
<td>-0.68</td>
<td>0.13</td>
<td>1.28</td>
<td>0.89</td>
<td>0.11</td>
</tr>
<tr>
<td>1.41</td>
<td>-0.61</td>
<td>0.21</td>
<td>1.23</td>
<td>0.95</td>
<td>0.15</td>
</tr>
<tr>
<td>1.20</td>
<td>-0.55</td>
<td>0.03</td>
<td>1.00</td>
<td>1.02</td>
<td>0.03</td>
</tr>
<tr>
<td>0.97</td>
<td>-0.48</td>
<td>0.22</td>
<td>0.77</td>
<td>1.09</td>
<td>0.16</td>
</tr>
<tr>
<td>0.58</td>
<td>-0.41</td>
<td>0.22</td>
<td>1.02</td>
<td>1.16</td>
<td>0.21</td>
</tr>
<tr>
<td>0.76</td>
<td>-0.34</td>
<td>0.26</td>
<td>1.10</td>
<td>1.23</td>
<td>0.10</td>
</tr>
<tr>
<td>1.15</td>
<td>-0.27</td>
<td>0.00</td>
<td>1.00</td>
<td>1.30</td>
<td>0.03</td>
</tr>
<tr>
<td>0.50</td>
<td>-0.20</td>
<td>0.06</td>
<td>0.80</td>
<td>1.36</td>
<td>0.09</td>
</tr>
<tr>
<td>1.05</td>
<td>-0.14</td>
<td>0.27</td>
<td>0.96</td>
<td>1.43</td>
<td>0.23</td>
</tr>
<tr>
<td>1.09</td>
<td>-0.07</td>
<td>0.26</td>
<td>0.88</td>
<td>1.50</td>
<td>0.14</td>
</tr>
<tr>
<td>1.00</td>
<td>0.00</td>
<td>0.27</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
statistics would clearly suggest that the item get flagged as well. In the fourth column, we again see evidence of distortion from the checks, but the fit statistics are not nearly as noteworthy. This is interesting. This is presumably due to the increased discrimination. This increased discrimination not only changes the fit statistics, but the violations detected on the low end of the ability spectrum disappear and only a bump on the right hand side remains.

In general, this figure suggests that one can probably simply “flag” an item using either the checks or the fit statistics.\(^8\) The checks, however, provide additional information above and beyond the simple flag. They indicate where an item is functioning poorly and this may lead to some natural fixes. Returning to item 14, in the second column it seems to suffer from guessing (note again that bump on the left). Perhaps this information could be used to fine tune the multiple choice responses. If an iterative approach to test design is being used, this information may be useful in better tailoring certain items to the Rasch model in subsequent version of the assessment. Item fit statistics could certainly be used in a similarly diagnostic approach, but they would typically require an a priori hypothesis about the nature of the problem (e.g., in the case of item 14, guessing is an issue) so that a fit statistic designed to assess the problem can be used. The checks, on the other hand, can be run without any prior hypothesizing.

As a brief digression, let us consider whether the fit statistics seem to be behaving as expected. Beginning with the mean-squares statistics, Wu and Adams (n.d.) suggests that these are related to the slope of the item characteristic curve.\(^9\) Can we observe this in Figure 4.6? In some cases, yes. Item 33, for instance, has reasonable mean-squares statistics in the third column (even though there is a touch of guessing) but not in the fourth column when the discrimination goes from 1 to 0.69. However, in other cases, the items seem to respond fairly strongly to guessing. For instance, look at item 2 going from column 3 to 4. Although the discrimination only changes by a small amount, the mean-squares statistics change fairly dramatically, apparently since the guessing parameter jumps

---

\(^8\) An approach that I have not explored, but which may be quite similar to the checks in terms of diagnostics, is item fit plots.

\(^9\) “The unweighted residual-based fit statistic provides a comparison of the slope of an observed (or empirical) item characteristic curve with the average slope of all item characteristics curves (ICCs) (i.e., average across all items)” (Wu & Adams, n.d., p. 16).
from 0 to 0.23. Based on these examples, one should be cautious about immediately concluding that discrimination was the issue based on merely knowing that an item had deviant mean-squares statistics.

Consider now the claim that the Q-index (Rost & von Davier, 1994) “is essentially a conjoint additivity test for a vector of (person or item) responses” (Karabatsos, 2000, p. 171). Evidence for this claim is mixed. Figure 4.7 plots this fit statistic as a function of the mean proportion of cells that had detected violations from the checks for each item. I use the same set of simulation iterations (with equal numbers of deviant guessing and discrimination items) as in Figure 4.6. Black dots are the purely Rasch items and red dots are those that are not purely Rasch. There are two clear clusters forming. Each cluster starts near (0, 0.25). The less deviant cluster, which contains the purely Rasch items, moves right and down from there. The more deviant cluster, moves up and right. After examining these cluster, it is not clear what is causing this separation. One natural hypothesis would be that these could be due to the guessing and discrimination deviations behaving differently with respect to the q-index, but this does not seem to be the case. There does seem to be some justification for Karabatsos’s claim for the items in the upward moving cluster, but more work is needed to say authoritatively how the fit statistic is related to the checks.

Finally, let us examine the claim (in which $H$ represents the omnibus scalability coefficients) that “a scale is considered weak if $0.3 \leq H < 0.4$, a scale is considered moderate if $0.4 \leq H < 0.5$, and a scale is considered strong if $H > 0.5$” (van der Ark, 2007, p. 4). How do these guidelines square with the simulation? The simulation based on 40 items showing each kind of deviancy (40 items with non-zero guessing parameter as well as 40 items with non-unity discrimination parameter) is a weak scale indeed, producing a statistic of only 0.16. However, the scale based on truly Rasch items only produces a statistic of 0.23. By the above criteria, this would not even rise to the level of a weak scale! This result seems to suggest that this statistic will be of only limited utility for use with large scale assessment.
Figure 4.7: Comparison of Checks to Item Q-index. Black dots represent Rasch items while red dots are items with some deviation from the Rasch model.
4.4 Difference Matrices

Turning away from the checks, I now focus on detecting scale distortions. One difficulty with thinking about scale distortions is that it is not necessarily positions on the scale that are most distorted but differences. As a single example, consider two individuals whose true values on a scale are both 500 where the standard error of measurement is 30. Due to item-side deviations from the Rasch model, they are estimated at being at 490 and 510 respectively. While these numbers may be reasonable estimates of the true ability, the difference between these students looks potentially meaningful when it should actually be null.

To focus on differences, consider the following object. For a single individual, consider the vector created by taking differences between their estimated ability and the estimated ability of every other individual. To keep quantities from cancelling, use the absolute value of each difference. This vector contains all available information at the student-level about differences across the scale. I refer to the full matrix of these quantities as difference matrices.

Denote a generic difference matrix by $\Delta$. Define $\Delta_{ij} = |\theta_i - \theta_j|$. Using the hypothetical estimates from above, we would have

$$
\begin{bmatrix}
|490 - 490| & |490 - 510| \\
|510 - 490| & |510 - 510|
\end{bmatrix}
= 
\begin{bmatrix}
0 & 10 \\
10 & 0
\end{bmatrix}
$$

The ability values can be estimates or, in the case of simulation studies, true values. Here I create difference matrices using both. One immediate difficulty with the difference matrices is that they provide too much information. In the case of the second simulation, the 15,000 students lead to $15,000 \times 15,000$ matrices. They must be summarized. Let $\hat{\Delta}$ represent the difference matrix formed by some ability estimates (both 3pl and Rasch estimates are used) and $\Delta^T$ be the difference matrix when the true abilities in the simulation are used. I then compute $\Delta' = \hat{\Delta} - \Delta^T$ and $\Delta'' = \hat{\Delta} / \Delta^T$ where $-$ and $/$ represent element-wise subtraction and division of the matrix elements. Finally, take medians over each column while excluding the diagonal elements. Medians are taken, as opposed to means, as small estimates tend to lead to especially large values in $\Delta''$. 
Since these objects are novel, Figure 4.8 was created to demonstrate how they behave. I created this figure by first sampling 100 points from a normal distribution, these are treated as the true values. Next, the value of each point was doubled so as to create a hypothetical “estimated” scale. It is then straightforward to create $\Delta'$ and $\Delta''$ as described and plot the medians of each column as a function of the true value. The uniform stretching leads to a very simple result for $\Delta''$ as the medians are just a horizontal line at 2. The effect on $\Delta'$ is slightly more complex, leading to the a shape similar to the graph of an absolute value function (note that the two sides each have slope of magnitude roughly 1). When the scale is compressed, this absolute value shape opens downward. If the estimated and true scale are the same, both lines are horizontal where the dashed line has an intercept at 1 and the solid the solid line an intercept at 0.

With the second simulation data, one more simplifying step is required. Rather than using the median of each column, I consider LOESS curves where the dependent variable is this median and the independent variable is the true ability estimate for each column/individual. These are shown in Figures 4.9 and 4.10. The solid lines in Figure 4.9 are centered around the horizontal line at 0 and are based on $\Delta'$. Beginning in the upper left, the solid red line indicate an asymmetric deviation at the low end of abilities for the 3pl estimates. This is due to the fact that the guessing parameters are not being fixed to 0 in the estimation of item parameters under the 3pl. As a consequence, estimates above 0 for the guessing parameter in the 3pl mean that certain correct responses due to ability are being mistaken as guesses. This leads to lower ability estimates for these respondees than are appropriate. The dashed lines in Figure 4.10 are centered around the horizontal line at 1 and are based on $\Delta''$. Starting in the upper left, the dashed lines seem to indicate minimal stretching (the line is near 1) although the stretching is not quite uniform (note the bump at the center). For comparison, the bottom right panel shows roughly even scale stretching for the 3pl abilities but seriously non-uniform scale compression for the Rasch abilities.

Looking down the first column of both Figures 4.9 and 4.10, the existence of guessing leads to peculiarities for the Rasch estimates. The Rasch estimates are producing differences that are much smaller than the true differences on the left hand side of the scale (see the solid lines). Since
Figure 4.8: Effect on $\Delta'$ and $\Delta''$ for case of uniform stretching. $\Delta'$ is solid, $\Delta''$ is dashed.
Figure 4.9: LOESS curves for $\Delta'$ when difference matrices of ability estimates (black-Rasch, red-3pl) are compared to true differences matrices.
Figure 4.10: LOESS curves for $\Delta''$ when difference matrices of ability estimates (black-Rasch, red-3pl) are compared to true differences matrices.
guessing is not being “controlled” for in the estimation of item parameters, this leads to a scale shrinkage for the low abilities. In contrast, the red lines suggest that the 3pl estimates are doing a decent job of scale preservation. The presence of differential discrimination manifests itself as a slight uptick in the solid red curve in the top right panel. In general, one can see a fairly large effect for all of the Rasch difference matrices along the main diagonal when compared to the truly Rasch case. Moreover, it is an asymmetric effect that tends to be more apparent at lower abilities. This is a critically important point as the asymmetry may lead to problems that would not occur if distortions were all symmetric.

The difference matrices have revealed some important facts about what each type of ability estimate does to the scale. In particular, the asymmetric effect noted for the Rasch case suggests that simply ignoring the lack of quantitative structure with given data may have consequences above and beyond extra estimation error. If estimates from a certain spectrum of the ability are more likely to be biased, this could lead to systematic errors. Such possibilities are considered in two simulations that I discuss later in this Chapter. At present, I turn now to a second version of the second simulation meant to address the afore-mentioned theoretical issues with the original version.

4.5 Second simulation Redux

This section discusses the impact of modifying a few assumptions from the above simulation on the ability scale. The first major change in this simulation is how abilities are sampled. Instead of sampling 15,000 abilities from the standard normal, I sampled forty-six abilities from a standard normal and then distribute these abilities to 15,000 examinees such that the resulting distribution of abilities are roughly normal.\footnote{I hypothesize here that ability is roughly normal as opposed to, say, uniform.} Next, I generate probabilities based on these abilities and the three-parameter IRF—in some cases, such as the baseline, this collapses to the Rasch model—that are used in turn to simulate response data. Finally, Rasch and 3pl abilities were estimated based on the response data. The second major change in this analysis is in judging the quality of
the resulting estimates. Instead of comparing estimates to true abilities, a perhaps unreasonable benchmark since any estimate is going to suffer from measurement error, they are compared to the Rasch estimates generated from truly Rasch data. I refer to these as the true-Rasch estimates and they are meant to serve as an ideal scenario in that all of the axioms are being met and interval scales are plausible.

Let’s begin with a look at the scatterplots for the case of truly Rasch underlying conditions, these are shown in Figure 4.11. Note that the Rasch and 3pl estimates are correlated at essentially unity while the estimates are both correlated with the true values at around 0.71. The Rasch estimates “stretch” the true ability scale slightly. The SD of true abilities is 0.34 while the SD of the Rasch abilities is 0.49. The 3pl abilities, on the other hand, have an SD of 1.4. That’s quite a bit of scale stretch! However, the stretching won’t necessarily be a problem so long as it is consistent across the scale (the result of a linear transformation).

To compare the ability estimates over each iteration of the simulation, Figure 4.12 shows the mean absolute difference between the estimated abilities and the Rasch estimated ability in the true-Rasch simulation. The x-axis shows the simulation index, which is the row number from Table 4.1. Row 1 is the true-Rasch simulation, note that there is no mean absolute difference for one of the black lines. That black line is the Rasch line. The horizontal black line is the mean absolute difference between the true-Rasch estimates and the true abilities. I plot it to provide a baseline for what kinds of distortions measurement error alone causes. The red line is the 3pl line. It is uniformly higher than the black line. What is rather surprising is how little both lines change over the course of the simulation. One can certainly detect a pattern in the red line (this pattern is showing the repeated increase in the number of items with guessing), but the black line is quite flat. It is not surprising that the Rasch estimates and the true-Rasch estimates are more similar to each other over the various conditions in the simulation than the true-Rasch and 3pl estimates, but it is surprising that the estimates are so consistent given the number of changes that occur between simulation 1 and 16.

From this perspective, the Rasch estimates are clearly preferable, in terms of how they
Figure 4.11: Scatterplots of abilities for secondary simulation in case of Rasch items

Figure 4.12: Mean absolute deviations between abilities and true-Rasch abilities. Black represents Rasch and red represents 3pl.
replicate the true-Rasch estimates which are as close as we can get to interval estimates in the presence of measurement error. That the more flexible 3pl model performs less ably here may seem surprising, but this actually harkens back to a point made by B. D. Wright (1997, p. 40): more parameters are not necessarily better. To further highlight the differences between the various scales in this scenario, I return again to the difference matrices introduced previously.

### 4.5.1 Difference Matrices, Revisited

The difference matrices introduced earlier can give us some sense for the global distortions that the interaction of the item-side deviations and the choice of scaling method produce. Figures 4.13 and 4.14 were created using a similar approach to the previous section but with one key difference. In this Figure, $\Delta'$ and $\Delta''$ are both defined as before with the exception that I now use $\hat{\Delta}_0$, the difference matrix formed using the true-Rasch estimates, in place of $\Delta^T$.

Compared to Figure 4.9 and 4.10, this simulation provides a more puzzling set of results. Consider first the Rasch estimates (black lines). In Figure 4.9, we saw that Rasch estimates tended to differ more extremely from the true values on the left hand side of the ability distribution. In contrast, Figure 4.13 does not have this asymmetry. The dotted black line in Figure 4.14 suggests that the compression is not entirely uniform, being more pronounced in the two regions between zero and the tails of the distribution. Although this is evidence of somewhat odd behavior from the Rasch estimates relative to the true-Rasch estimates, the 3pl estimates are much more erratic.

The 3pl lines (in red) have two distinct properties (especially outside of the top left panel). First, the dotted line in Figure 4.14 for $\Delta''$ suggests that there is a severely non-uniform scale expansion occurring (expansion since it is centered above unity and non-uniform since it is not flat). The line for $\Delta'$ (differences) in Figure 4.13 tends to be relatively flat and above 0. I am not sure what kind of scale transformation would produce such an effect. One possible explanation would be that there is some kind of scale translation occurring. To test this, I regressed 3pl and Rasch abilities on the true-Rasch abilities (and an intercept). In the 3pl case (as well as the Rasch case), the estimated intercept averaged over all 16 simulation conditions was under 0.01 in
Figure 4.13: LOESS curves for $\Delta'$ when difference matrices of ability estimates (black-Rasch, red-3pl) are compared to true-Rasch differences matrices.
Figure 4.14: LOESS curves for $\Delta''$ when difference matrices of ability estimates (black-Rasch, red-3pl) are compared to true-Rasch differences matrices.
magnitude so this does not seem to be a likely answer.\footnote{It’s interesting to note that when both the Rasch and 3pl ability estimates are regressed on the true abilities and an intercept, the average estimated intercept is quite large, -0.1 for the Rasch estimates and -0.3 for the 3pl abilities. Furthermore, while the average Rasch slope is almost flat (0.96), the average 3pl slope is quite large (2.97). Since the true abilities were not used in the construction of either $\Delta'$ or $\Delta''$, this can’t explain the phenomena.} These regressions do offer additional support for the scale compression of the Rasch estimates (an average slope of 0.5) and the scale expansion of the 3pl estimates (an average slope of 1.55). While further exploration would be needed to understand the exact distortions occurring in the 3pl ability estimates, it suffices to say that the scale of the 3pl estimates seems to be decidedly non-interval.

The point of this redux was to highlight the influence of a few assumptions made in the second simulation. Most of the above discussion has focused on the comparison of estimates to estimates rather than true abilities. The effects are quite pronounced. Future research based on simulation of Rasch models should be aware of these differences. I have said less about the effect of sampling at only a small set of abilities. This is due in part to the fact that this change’s effects on the simulation are not immediately obvious. But, more importantly, I demonstrate below that sum scores are a pretty poor indicator of ability. This being the case, the differences in simulation that I’ve discussed may have relatively minor implications compared to the noise induced by the fact that the sum scores are such poor indicators.

### 4.5.2 Quality of sum scores as indicators of ability

In each iterate of the simulation, I computed a weighted average of the standard deviations of true abilities within each sum score (weighted by the number of respondents with a given sum score) and compare this to the standard deviation across the entire set of true abilities. For the fully Rasch iteration, this weighted proportion is 68\% (so, on this weighted average, 68\% of the variability across all true scores exists within each sum score). It rises to a mean of 70\% across all the iterations, but these are of less interest since the sum score is a sufficient statistic for ability only in the Rasch case. Despite this, a single sum score contains a great degree of variability in terms of true scale scores. Another perspective is to consider the number of true scores contained within each sum score grouping. Taking another weighted average, nearly 25 separate true scores...
are represented within each ability grouping based on sum score in the Rasch iterate. The sum scores is a rather noisy estimate of student ability. Even if the sum scores are suboptimal ordinal indicators, the success of the checks in determining distortions in the simulations discussed here means that they are not worthless. It may still be worth considering alternatives. One possible alternative may be to validate vertical scales in experimental settings where students are exposed to different levels of instruction.\footnote{As there is no uncertainty in the level of instruction, the students can be unequivocally ordered.}

4.6 Effect of Distortions

In the following two sections, I attempt to demonstrate how wrongly presuming that a scale is interval can lead to incorrect inferences about the nature of the world. In the first demonstration, I imagine that some educational intervention has been implemented (in a manner like a randomized controlled trial) and analyze the estimation of the intervention’s effect (in a pre/post-test scenario) when items do and do not follow the Rasch model. In the second, more elaborate, demonstration, I describe a method of clustering students into schools via simulation. Based on this clustering, I try to tease out the impact of distortions at various points in the scale.

4.6.1 Treatment Simulation

I begin by describing a relatively simple method of studying the impact of the erroneous assumption of interval scales. In this simulation, I took 6 relatively simple steps meant to simulate the estimation of a treatment effect in a pre/post-test situation.\footnote{I'd like to thank Derek Briggs for suggesting this idea of estimating a treatment effect.} 

(1) Simulate 10,000 known “true” abilities from the standard normal and a set of 45 item parameters. The item parameters were one of two varieties: either truly Rasch with difficulties ranging from -1.5 to 1.5 evenly or the same difficulties but with guessing and discrimination parameters randomly sampled from uniforms distributions on [0, 3] and...
97

[.5, 1.5] respectively. These are the same distributions used to generate item parameters as in the second simulation.

(2) Simulate responses based on both sets of items and then estimate Rasch abilities based on those responses. Recall that the estimates based on the items which conform to the Rasch model will produce a scale where the data meets the axioms of conjoint measurement while the estimated abilities based on the alternative set of items will not. These are the pre-test abilities.

(3) Randomly assign roughly 1/3 of the 10,000 to a treatment group and add 0.2 to their true abilities. Note that this implies that the effect is constant across all students. While a homogeneous treatment effect for all students may be implausible, interest will be on recovering the mean pre- to post-test difference so the individual-level effects will not be a focus. The size of the effect is motivated by findings from the Tennessee STAR experiment (for example, see effect size for Word Study Skills for all students in grade 2, Finn & Achilles, 1999, Table 1.1), a reasonably large treatment effect for a major educational intervention.

(4) Simulate a new set of responses using the updated true abilities (same as original for control, plus 0.2 for the treatment) but the same item parameters as step (2). Using the Rasch model, estimate the post-test abilities based on these responses.

(5) Compute the effect of the treatment. Since linear scale transformation should be permissible if the underlying variable has interval properties, I ensure comparability of the various ability estimates by standardizing the pre- and post-test scores separately.

(6) Perform the above steps 100 times.

Although the 3pl item parameters here are simulated, they are actually fairly similar to what one would see in practice with assessment data that gets scaled with the 3pl model. To demonstrate this, consider the following statistics for 3pl item parameters from 6 grades and 6 years of CSAP Reading data (36 sets of item parameters). The mean 25th and 75th quantiles of the distribution of
item discriminations (over the 36 sets of data from the CSAP) were 0.67 and 1.13. Expected 25th and 75 quantiles from a Uniform distribution on 0.5 to 1.5 are 0.75 and 1.25. Similarly, the mean 75th quantile from the distribution of guessing parameters in the CSAP data was 0.25 while the expected value here would be 0.23. Since the deviations that have been introduced are consistent with what we observed on data that is thought to require the 3pl, the scale distortions should be similar to what we would observe if we mistakenly scaled such non-Rasch data with the Rasch model.

Across the 100 iterations, there is evidence that if the treatment and control means are computed from a test with a non-interval scale then the result is a slight underestimation of the true effect. Looking at the mean post- to pre-test differences between the treatment and control, the mean over the 100 iterations with the Rasch items was 0.21. With the non-Rasch items, it was 0.17. This is a pretty small underestimate that would unlikely lead to major mistakes in terms of how the treatment is evaluated or applied. However, there is a more intriguing problem lurking under the surface. When I regress the score differences on a dummy variable indicating treatment interacted with the prior test score, the average interaction is much larger in the case of non-Rasch items (remember that the treatment effect was constant so there should be no interaction with ability).

Over the 100 iterations, this interaction averaged 0.002 for the Rasch items and 0.018 for non-Rasch items, nearly 10 times higher. Furthermore, the effect is significant at the 5% level in 35 of the 100 cases for the non-Rasch items and only 10 of the cases for the truly Rasch items. The interaction is illustrated in Figure 4.15 in which I show the difference between truly Rasch and non-Rasch items (remember that the abilities are the same) on the slope of the least squares lines for one of the 100 iterations. At the 25th percentile of the pre-test distribution of estimated abilities, the estimated difference in differences would be 0.2. For those at the 75th percentile, this difference in differences is 0.25. This is an interesting result, but still probably not enough considering the difference at the 25th and 75 percentiles, to lead to major errors in judgment.
Figure 4.15: Regression lines from regressing post- to pre-test difference on the pre-test score interacted with a dummy variable indicating treatment. The black line are the predicted differences for the control group and the red line represents the fitted differences for the treatment group. At top, the data is purely Rasch and the constant treatment effect is roughly recovered (note that the interaction p-value is relatively large. At bottom, the items do not conform to the Rasch model and this suggests, incorrectly, that lower ability groups receive less of an effect from the intervention than higher ability groups.
To really stress the scale, the simulation was rerun with more extreme item-side deviations. The discriminations for the 45 items were drawn from a uniform distribution on 0.25 to 2.5 and the guessing parameters were drawn from a uniform distribution on 0 to 0.6. These are far larger deviations from the Rasch model than any of the sets of CSAP item parameters considered earlier (although, perversely, not that far from 3pl estimates based on a purportedly Rasch set of responses that I consider in the following Chapter). In this simulation, the mean interaction over the 100 runs was 0.00 in the true Rasch case and 0.02 in the non-Rasch case. Although more of the 100 iterations here showed a significant interaction in the non-Rasch case (48 of the 100 iterations), this is still a fairly small interaction effect. Taken as a whole, the evidence suggests that the Rasch scaling method is fairly robust to the kinds of distortions that have been induced here.

4.6.2 Clustering Simulation

The next simulation is meant to highlight the kinds of errors one might make if data inappropriately scaled with the Rasch model is aggregated to a higher level (such as a school). It is again designed to demonstrate the danger of assuming quantitative structure for some underlying latent variable where none exists. This simulation is based upon an analysis of the school-level variation of student abilities. This is motivated, as I explain below, by what was observed in the difference matrices. This clustering algorithm works by assigning students to schools via a draw from a multinomial distribution.

To assign students to schools, my approach is to consider a mixture of normals where each normal is associated with a certain school. For a student with a certain ability, the probability of attending a given school is proportional to the density of that school’s associated distribution at the student’s ability. The motivation here is that schools typically have more students with more similar abilities than the district in which the school is situated (within-school variances of student abilities are smaller than within-district variances). To set up the mixture models, I first sampled the parameters controlling each normal distribution. Let $\mu_i \sim N[0, 1]$ and $\sigma_i^2 \sim \text{Unif}[1, k]$ where $i \in \{1, \ldots, N\}$ and $N$ is the number of schools in the simulation. I set $N = 50$. Given the starting
point of a population of 15,000 students this is roughly 300 students per school. Note that “school” here is actually a slight misnomer as I was only considering a single group of students, so a better comparison is to a single grade within a school. For the simulation discussed here, \( k \) was set equal to 5. The choice of \( k \) controls how homogeneous each school is relative to the overall population of students. Larger values of \( k \) correspond with higher within-school variances and lower between-school variances. In additional simulations not discussed here, I varied \( k \) to make sure that the simulation behaved reasonably (e.g., the variance in school means was larger for smaller values of \( k \)).

I then drew 15,000 student abilities from the standard normal.\(^{14}\) For student \( j \) with true ability \( \theta_j \) I computed

\[
d_{ij} = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{(\theta_j - \mu_i)^2}{2\sigma_i^2}\right)
\]

for each \( i \). This is simply the density of \( N[\mu_i, \sigma_i^2] \) at \( \theta_j \). A multinomial distribution was then created that gave the probability of student \( j \) attending school \( i \), \( P_{ij} \), by setting

\[
P_{ij} = \frac{d_{ij}}{\sum_i d_{ij}}.
\]

A single draw from this distribution leads to a school assignment for student \( j \) and this process is repeated over all students.

Remember that the impetus is to simulate the kind of between-school heterogeneity that exists in real educational settings. The amount of within-school variance compared to between-school variance may vary substantially by setting (e.g., rural versus urban). I am able to control this through choice of \( k \). How “lifelike” is the approach using the values specified here? To answer this question, I compared between versus within school variances using F-statistics based on a regression of student scores on school IDs. As a baseline, I used state-wide assessment data from Colorado. Between grades 6 through 9, Colorado had roughly 54,000 students with between 90 and 130 students per grade (fewer than the 300 I use here). The state produced F-statistics between 25 and 33. To replicate these F-statistics in the simulated clustering, I experimented with various

\(^{14}\) I in fact use the abilities and estimates from the second simulation.
values of \( k \) before finally choosing \( k = 5 \). With this value, the F-statistics were between 27 and 33.\footnote{Higher values of \( k \) resulted in lower F-statistics and smaller values of \( k \) in higher F-statistics.} Although my simulated data has fewer schools and students overall, the within-school to between-school variance ratio is about the same so the below results should be similar to what we’d expect to see in the real world.

At this point in the process, students are clustered within schools based on their true abilities. The naive assumption made in all cases is that the data is truly Rasch and that the Rasch estimates are on an interval scale. The estimated abilities in fact come from the second simulation iterations where the number of item distortions varies (in particular, the number of non-zero guessing parameters and non-unity discrimination parameters varies between 0, 10, 20, and 40 items). The question of interest is then: What is the effect of ignoring these distortions? Recall the results from the analyses of the difference matrices. In the presence of substantial item-side deviations, the Rasch estimates compressed the scale non-uniformly. In particular, the compression was more extreme for lower abilities. This suggests that the variability within schools could be mis-estimated as a function of where the schools are on the scale since the differences between a student’s ability and the mean school ability may be compressed different amounts. In particular, lower ability schools should have the variability of their students’ abilities under-estimated.

Figure 4.16 is an analysis of the school standard deviations as a function of school mean that is meant to determine whether the conjecture made above is actually correct. I first standardized all true and Rasch scores and then compared the within-school SD to the school mean in Figure 4.16. The black dots are the true means and SDs. The black curve is a LOESS curve fit through the black points. The red dots and curves are the same but for the standardized Rasch estimates. From left to right within a row there are increasing deviations in the discrimination parameters and from top to bottom within a column there are increasing deviations for the guessing parameters. Starting in the upper left and moving across the top row (where more items are deviant in terms of their discrimination but none have a non-zero guessing asymptote), the two LOESS curves track each other quite well. However, as you move down columns, you can generally see the red line be
Figure 4.16: Scatterplots comparing the abilities on the x-axis (standardized true abilities in black, standardized Rasch abilities in red) to the standard deviation of each quantity on the y-axis. The similarly colored curves are LOESS curve fitted through each set of points. Each dot represents a distinct school mean ($N = 50$).
below the black on the left and above the black on the right. This means that relative to the true variation of abilities within a school, the Rasch estimates are underestimating the ability variance for low-ability schools and over-estimating the variance for high-ability schools.

Since columns represent increasing guessing but no changes in discrimination, it is clearly guessing rather than changes in the discrimination that is driving the the divergence between true and Rasch ability estimates. This is not surprising given the interpretation of the difference matrices above. Moreover, it is intuitive in the following way. Changes in discrimination are perhaps a problem, but they are symmetric about the difficulty of a given item so should impact the students at each end of the ability distribution in a roughly similar manner. This results in situations like Figure 4.16, where distortions are going to be a function of scale position. If the bottom row of Figure 4.16 is taken to represent reality (instead of the correct picture in the upper-left), we would mis-estimate the nature of ability clustering within these schools although perhaps not in all that serious a manner (differences between true and estimated SD of school abilities based on the LOESS curves never seem to go above 0.2).

4.7 Discussion

In the discussion, I highlight a few of the main findings of this chapter and then specifically address the three research questions posed at the beginning of the chapter. The first set of simulations were designed to analyze the statistical properties of the modified checks. In a simulation where the underlying data is generated by a Rasch model, the checks detected distortions—meaning violations of the cancellation axioms of ACM—at a relatively low rate (1.3% in the size analysis and 2% in the first iteration of the second simulation) compared to the choice of a 95% credible region from the posterior. However I also demonstrated that this could be controlled as a function of the size of the posterior credible region that one chooses to work with. Not surprisingly (given that there are fewer observations at the extremes), distortions are over-detected at the extreme sum scores.

When I modified a single item to be inconsistent with the Rasch model, by either varying the
discrimination or giving the item a lower bound on the probability of a correct answer, the checks were able to detect the resultant distortions although there is naturally some sensitivity to the size of the deviation. Interestingly, detecting distortions depended on factors besides simply the item parameter that had been changed as the ability to detect distortions depended in some cases upon the value of the item difficulty. For example, it was difficult to detect that easy items may have guessing occurring primarily because the ease of the item ensured that few students would be in the region of the ability scale where guessing is causing distortions. I relied primarily on visual inspection (of graphics showing the the proportion of detected violations as a function of ability) in this endeavor. This was due to the fact that a specific deviation from the Rasch model for an item does not necessarily impact all of the sum score cells for that item.

I intended the second simulation to be a more realistic assessment of the checks. Comparing mean detection rates across items, the checks did seem to detect items with deviations from the Rasch model with one caveat. When a large number of the items had item-side deviations, the checks detected violations at relatively high rates for both the truly Rasch and deviant items. This is perhaps not that serious an issue. Consider the case from Table 4.1 with 20 items showing each type of deviation. The weighted mean for the “ok” items is half that of the weighted mean of the discrimination item. The weighted mean of the guessing items is only slightly above that of the Rasch items, but this is part of the general inability of the checks to distinguish items with non-zero guessing parameters. In the case of 40 deviant items of each type, the weighted mean for the Rasch items is too high, but there are so few of them that the instrument probably isn’t worth trying to reformat as a Rasch instrument anyways. Generally speaking, the checks seem quite able to distinguish items with non-unity discriminations but have less ability to distinguish items with non-zero guessing parameters. This is somewhat disappointing since the guessing parameters seem to be the main cause of non-interval scale distortions (of the type discussed in the treatment and clustering simulations).

The second simulation was based on drawing a large number of unique abilities from the standard normal distribution. However, it may be unrealistic to expect the Rasch model to be able
to preserve such fine-grained differences when there are only $n + 1$ distinct scores for an assessment with $n$ items. So, in a redux of the second simulation, I examined differences between the Rasch and 3pl scales when a much smaller set of abilities was drawn from the standard normal and then a large number of people was assigned to each of these abilities. Furthermore, subsequent comparisons were made not between estimates and the underlying “truth” of the simulation, but between estimates that occurred when the items were not purely Rasch and the “true-Rasch” estimates that occurred when the items were truly Rasch. Relative to the level of underlying measurement error, the mean-absolute deviations between the Rasch and true-Rasch estimates in any configuration of the simulation were relatively small. The 3pl estimates exhibited much larger mean-absolute deviations relative to the true-Rasch estimates. When I compared difference matrices based on estimates to the difference matrix based on the true-Rasch estimates, I noted that the 3pl estimates behaved rather oddly and that the Rasch estimates tended to compact the scale. I also noted the fact that the sum scores were a surprisingly poor indicator of true ability.

Let me now return to the three research questions stated at the beginning of this Chapter. The first question is whether the checks are reliable statistical tools. Based on the simulation evidence presented here, it is safe to conclude that the checks are a mostly reliable method of detecting violations in the cancellation axioms of ACM. More work would be needed to diagnose how they respond to deviations from the Rasch model not considered here, but at least in the item-side deviations I have considered they seem to behave as one would expect. The one possible caveat might be that they are poorer indicators of items that have problems with guessing than items with faulty discriminations.

The second research question was how the checks can be distinguished from fit statistics. The work here suggests that items are likely to be “flagged” using either the checks or the fit statistics but that the checks may give additional diagnostic information (like suggesting that guessing is occurring in the above example) that could be quite useful. Alongside this diagnostic information, I can think of two other potentially useful attributes of the checks. First, they don’t have the same dependence on sample size. Hence, one does not need to worry about identifying critical values
such as the 0.8 and 1.2 that are frequently suggested for mean-square statistics (these are actually conservative, both Bond and Fox (2001, p. 129) and Wilson (2005, p. 178) suggest 0.75 and 1.3) but probably inadequate for most samples of more than a few hundred people. Second, they could perhaps make people more critical of model fit and the quality of the scale that results from fitting the Rasch model since they can be obtained prior to Rasch estimation (the scalability coefficient also shares this trait). Once Rasch estimates are in hand, the natural inclination is going to be to use them as though they are interval. Being able to analyze beforehand whether the equal-interval assumption makes any sense could be beneficial. My logic here mirrors an argument from Rubin (2008) in a causal inference setting. Although it sounds fanciful at first, I think that we perhaps underestimate the role of human nature in statistics. Merely making people ask questions about the quality of the Rasch estimates that they are likely to get before they get them could lead to more thoughtful uses of large-scale assessment data.

The third question was what are the impacts of scale distortions? There have been three explorations of this: the difference matrices, the experiment, and the clustering simulation. Starting with the difference matrices, these objects pinpointed the scale distortions caused by certain item-side deviations. For example, in the redux of the second simulation, it was clear that the Rasch and 3pl estimates “stretch” the scale in different ways. Although their use here was always relative to true data, the difference matrices could potentially allow for the analysis of distortions in empirical scales. In any event, these analyses suggest that we should have certain expectations from empirical data. Imagine a context in which the Rasch model is to be used for scaling. If there is reason to suspect that students are guessing on a substantial chunk of the items (perhaps as indicated by the checks), then this is likely to lead to distortions of the kind seen on the lower right of Figure 4.9 and 4.10. Discrimination seems like a less serious problem in terms of scale distortions. Highlighting the effect of such distortions was the goal of the experiment and clustering simulations.

---

16 As they are used here, it is true that the difference matrices are only useful when truth is known. However, I am not convinced that further study of these objects couldn’t lead to their utility in empirical work. For example, imagine simulating data such that one understands the underlying deviations from the Rasch model. Using just estimates, one could produce difference matrices and examine the difference matrix as a function of the estimated abilities. If certain patterns could be identified as corresponding to certain types of deviations, then this could potentially be applied to empirical data when truth was not known.
In the experiment related to Figure 4.15, there seemed to be only a small penalty in estimating the treatment effect when one incorrectly assumed the scale was interval. Another possibility was that one may be lead to believe that the treatment effect was actually larger for higher ability students when in fact it was constant for students of all abilities. However, even here the effect was only minor. The effect of scale distortions on this simulation were fairly minimal. Similarly, in the clustering example, an incorrect assumption of linearity would lead to systematic under- and over-estimates of school ability groupings. However, it’s not clear how serious such systematic error would be.

In general, it seems quite difficult to “find” the effect of scale distortions. Although the difference matrices gave clear indications that in some instances the item-side deviations were leading to scale distortions that were not simply linear transformations, these distortions didn’t seem to affect most aggregated quantities. In the above, I only mentioned quantities that showed some differences but I also attempted to find the effects of differences in other quantities. For example, using the means from the simulation, I both ranked schools and computed an analogue of percent proficient (I arbitrarily used the origin as the proficiency threshold). In both cases, there was very little difference in the results whether I used the true or estimated school means, even when the estimated school means were computed using estimates based on the worst-case simulations. What both of these cases have in common, however, is that they only necessitate ordinal data. They do not require interval data. Thus, the failure to see differences in these cases should not be surprising. However, they are common examples of the ways that people use test score data.

Let me offer one final comment on the effect of scale distortions. To use the language of Donald Rumsfeld (February 12, 2002), the former US Secretary of Defense, measurement error is a known unknown (although one can make assumptions that make it a “known known”; indeed, that is the goal of most psychometric models) and scale distortions caused by axiom violations are unknown unknowns. A great deal of time and effort goes into understanding and minimizing measurement error by the psychometrics field. Far less goes into understanding the consequences
of non-linear scale distortions. The results presented here suggest that this ignorance may not have deleterious consequences. Rather than simply hoping that this continues to be the case, it seems more prudent to investigate this issue in greater detail.
Chapter 5

Empirical Demonstration

This chapter is meant to address the third research question, whether empirical datasets are consistent with the measurement axioms ([Q3] from Chapter 1). I focus on two assessments that differ with respect to several key design features but are each scaled with the Rasch model. I demonstrate that the data from one assessments is coming pretty close to meeting the measurement axioms while data from the other has large numbers of violations. However, even the assessment that meets the axioms does not seem to have been generated by the Rasch model.

5.1 Lexile Scale

MetaMetrics has incorporated an auspicious, from the perspective of ACM, design feature into their Lexile scale for reading comprehension that allows for the ranking of items in terms of difficulty prior to exposing the items to a given group of students. At the heart of this scale is a theory of reading difficulty (Stenner et al., 1983) that takes into account sentence length and word difficulty. Through this theory (and previous data collection), they are able to predict Rasch difficulties before a group of students is exposed to a given assessment. This allows for the manipulation of item difficulty. Recall that conjoint measurement requires the ability to manipulate both entities, in testing this would be ability and difficulty, simultaneously. With the Lexile scale, difficulty can be manipulated since items can be ordered by these predicted difficulties before they are administered. However, manipulation of abilities is not yet feasible so although their approach is an improvement it does not yet meet all of the requirements of conjoint measurement.
After examining data generated by the Lexile scale, Kyngdon (2011) claims there are indications (but stops short of absolute confirmation) that the Lexile scale is interval.\footnote{The relevant passage reads as follows: “On the basis of these results it can be concluded that the difficulty of reading items, as conceived of in the Lexile theory, and the reading ability of persons are quantitative. A comprehensive test of the Lexile theory, however, was not the goal of the current example and therefore any judgement concerning the descriptive adequacy of this theory is premature.” (Kyngdon, 2011, p. 488).} The strengths of the Lexile scale have led to direct comparisons with the measurement of temperature. For example: “Measurements for persons and text are now reportable in Lexiles, which are similar to the degree calibrations on a thermometer” (Stenner, 1996, n.p.).\footnote{The title of Stenner (1996), “Measuring reading comprehension with the Lexile framework”, is also meaningful given the philosophical stance of the author. He is not implying squishy psychometric “measurement” but something analogous to what occurs in the physical sciences.} Although I want to be clear in stating that the Lexile scale is an extremely interesting example of psychometric measurement and a potential starting point for developing interval scales, the evidence I will present indicates that the Lexile scale does not yet have equal interval properties that would make it analogous to the degrees on a thermometer.

5.1.1 Background on Lexile framework

The Lexile framework is a theory for reading comprehension based on text difficulty (Stenner et al., 1983). This theory posits that reading comprehension is a function of quantitative features of the text. The first such feature is sentence length. This is straightforward, simply the number of words in a sentence. The second feature is slightly more difficult to quantify. Words in a given passage are given a frequency index, a numbered formed on the basis of counting their occurrences in a standardized corpus (such as Carroll, Davies, & Richman, 1971). Using previously collected data, estimated Rasch difficulties were regressed on these two factors (for the word frequency, the mean of the logarithm of the frequencies is used). When new items are developed, their difficulties can be estimated a priori by using the characteristics of the text and these regression coefficients. Over 15 years of work with this system, this relationship between text difficulty and the Rasch difficulty coefficients has been shown to be quite sound (Stenner, Stone, & Burdick, 2011).

MetaMetrics has operationalized this concept as the Lexile scale. From the classical perspec-
tive on measurement, this scale is quite interesting since it has a base unit: “the unit of the Lexile scale, represented by an 'L', is defined as one-thousandth of the difference in difficulty between a sample of basal primer texts and Grolier’s (1986) Encyclopedia” (Kyngdon, 2011, p. 485, emphasis in original). Students are placed onto the Lexile scale via their performance on the “cloze” items that MetaMetrics has developed. These items contain a passage with the final word omitted. The respondent must choose the correct closing word to each passage (they must “cloze” the passage) from a set of responses. On the one hand, the design of the assessment is such that items can be ordered in difficulty before administration. This would allow for the possibility of creating data specifically designed to test the conjoint axioms, although as Kyngdon (2011) notes, the data used here was not designed with this goal in mind. On the other hand, while the Lexile framework does provide a method for estimating Rasch difficulty prior to item use, there is nothing in either the theory or the response format to ensure that guessing does not occur or that all items have equivalent discriminations. Just because sentence length and complexity predict difficulty, this does nothing to ensure that the items otherwise conform to the Rasch model. Since the assessment is based on a theory of reading comprehension rather than considerations of model fit, this could be a problem.

5.1.2 Data

I use the data collected in Duval, NC that was also used in Kyngdon (2011). There are 9638 4th graders with complete response strings and 54 items. Like Kyngdon, I focus on sum scores between 13 and 52 (inclusive) which were those sum scores that had at least 45 respondents. Kyngdon used a slightly larger group of respondents, 9708 students. Before moving on to an analysis based on large-scale sampling of 3-matrices, I start with the more modest goal of replicating the results in Kyngdon (2011), which used only 3 sum score groups of students and 4 items. The 3 sum scores (33, 34, and 35 correct out of 54 items) capture only a small amount of the overall sum

---

4 The difference in sample size is probably based on his potential inclusion of students with missing responses. It is unclear how he grouped those that had omitted responses to certain items. This is problematic since sum scores based on different number of total item exposures are not directly comparable.
score variability. The items were 52, 47, 34, and 16 (difficulties of 1080L, 970L, 800L, and 640L respectively).

5.1.3 Replication

The initial attempt at replicating the results of Kyngdon (2011) was a success. I found numerous violations related to items 47 and 52 when conjoint matrices were ordered by the previously determined Lexile difficulties. Kyngdon suggests reversing the order of these items. He justifies this procedure within the context of the Lexile framework by noting that the difference between the difficulties is within the error of measurement for the Lexile items. When I reversed the ordering of these two items, I found (like Kyngdon) that the axioms seem to hold in the $3 \times 4$ conjoint matrix with one slight caveat. I detected a distortion for item 16 (sum score 33). Kyngdon detected something similar, having this cell be nearly flagged in several cases, so this doesn’t seem to be a substantive difference between our results.

The re-ordering of the items by Kyngdon (2011) is important since it is what allows him to say “on the basis of these results it can be concluded that the difficulty of reading items, as conceived of in the Lexile theory, and the reading ability of persons are quantitative” (p. 488). Kyngdon justifies the re-ordering as follows:

Assuming the theory is true, Lexile item difficulty values have a systematic measurement error of 170L (Stenner et al., 2006). Item 52 is only 110L more difficult than item 47, so it is plausible that the independence axiom failed due to systematic measurement error. Permutation of columns 1 and 2 in Table 2 enabled a test of this hypothesis. The results are presented in Table 3. The independence axiom was supported. (p. 10)

While I agree with his reasoning within the context of his analysis, this approach is inconsistent with my repeated sampling approach. Consider the puzzle that would be encountered when working with the data matrix. If a single 3-matrix doesn’t support cancellation given the initial ordering but does under a modified ordering, should this modified ordering be extended to all of the cells for these items? It must for the procedure to be coherent (item orderings must be consistent across the ability scale), but this would require that all previous tests using these cells be repeated.
Since the following section conclusively demonstrates the superiority of the repeated sampling approach, modified orderings are not a viable approach outside of limited examples along the lines of Kyngdon’s.\footnote{There is a subtle point to be made here. Kyngdon (2011) is actually checking for independence in the entire $4 \times 3$ matrix in a single check. If this approach were viable, then re-ordering might also be viable. However, the difficulties related to the “mesh” that I have already explained mean single cancellation checks can only occur in modestly sized matrices. Since it is also important to focus on the entire data matrix rather than a mere subset, this means checking for single cancellation within 3-matrices and my reasoning about the difficulties regarding the re-ordering would then follow.}

### 5.1.4 Results

After the success in the initial replication, I looked for distortions in a large sample of 3-matrices. Note that I do not switch column orderings in individual checks since this would need to be done across the entire response matrix. Looking at Figure 5.1, results are not at all promising for the Lexile scale as far too many distortions are being detected here. It is important to note that the items are ordered not by sum scores but by the Lexile difficulties established via the theorized difficulty of the reading passage. This is an important point. The claim being made by Karabatsos is that the scale produced using these item difficulties (as opposed to item difficulties estimated in the standard way) is interval. To evaluate this claim, I focus on the ordering imposed by the theorized difficulties. There is a small pinch in the middle where items are showing distortions at slightly reduced rates, but the overall violation detection rate is entirely too high. Compared to the rates of 2\% for truly Rasch data (or even under 10\% in all iterations of the second simulation, see Table 4.1), 29\% of the checks here are reporting violations.\footnote{When the items are ordered by their sum scores (as opposed to their theoretical difficulty based on reading complexity), the percentage of violations falls to 11\%. While this is still far too high, it is certainly an improvement.} The conclusion here is obvious: the data does not seem consistent with the axioms of additive conjoint measurement. A great deal of skepticism that this data could be used to create an interval scale seems warranted.

**Fit statistics and the checks** The fact that the data are inconsistent with the cancellation axioms of ACM implies that it is impossible that the data are consistent with the Rasch model. This is supported by the fit statistics.\footnote{Note that the fit statistics used here are those I computed based on my estimates of the Rasch difficulty, not the official Lexile ones. Using the Lexile difficulties (those predicted by sentence length and difficulty) rather than those} I present the fit statistics both graphically and in tabular
Figure 5.1: Proportions of distortions based on Lexile ordering for Duval data.
Figure 5.2 gives a visual indication of fit while Table 5.1 contains all the information on the items (checks, item fit statistics, and 3pl parameters) discussed in this section. Figure 5.2 contains four panels, one for each fit statistic. The density curve in each panel shows the density of the fit statistics for the Duval data. The solid vertical lines are the mean and plus/minus one SD computed over each fit statistic for all items (the mean is the thickest of the three). In contrast, consider now the dotted lines. The heavier dotted line is the mean and the lighter dotted lines are plus/minus one SD for fit statistics computed in the true Rasch simulation from the prior chapter. The variability in the fit statistics when the Rasch model is the true model is far smaller than what is observed in the case of the Duval data. Although the sample size for the Duval data is slightly smaller than that for the simulation from the prior chapter (nearly 10,000 compared to 15,000), it is implausible that this difference could be driving the results shown in Figure 5.2.

Rather than consider the distribution of fit statistics, it is more typical to merely “flag” items based on their values. Here I focus on the mean-square statistics. As one can see in Table 5.1, using the standard 0.8 and 1.2 critical values (see Bond and Fox (2001) or Wilson (2005) for more liberal recommendations) for the mean-squares statistics, one would flag only 3 items as poor fitting based on outfit and 21 based on infit. However, Table 1 in Wu and Adams (n.d.) states that for 1000 persons and 40 items, the correct critical values for the mean-squares statistics are 0.9 and 1.1. Since 0.8 and 1.2 were far too lenient, items that may have misfit were not flagged. With critical values of 0.9 and 1.1, 36 and 28 items are flagged with the outfit and infit statistics respectively. Since there are nearly 10,000 people in the data, even 0.9 and 1.1 are too lenient as they were based on simulations containing only 1000 respondents. If one uses 0.95 and 1.05, 47 and 38 items are flagged by the outfit and infit statistics respectively.

Based on flagging items, we see that the estimated based on the data would presumably make the fit statistics even more deviant. Since those pre-determined Rasch parameters are less optimal than the estimated ones under the estimation criteria, the resulting residuals using the pre-determine parameters will be larger. Since many of the item statistics that I use are based on the residuals, this will cause the fit statistics to be farther from their expected value under the null hypothesis of model fit. The scalability coefficient does not depend upon item parameters, so those fit statistics would be unaffected by this switch. The choice of 0.95 and 1.05 is arbitrary, based solely on the fact that 0.9 and 1.1 are still too lenient. One could simulate exact thresholds for the fit statistics based on the number of items and the number of people (I take this approach in the next section), but I do not do so here since even with the lax thresholds of 0.9 and 1.1 over half of the items are being flagged, suggesting that the data is not Rasch.
Table 5.1: Item indicators for Duval data

<table>
<thead>
<tr>
<th>Item</th>
<th>% Checks</th>
<th>3pl Disc</th>
<th>3pl Diff</th>
<th>3pl Guess</th>
<th>OMS</th>
<th>IMS</th>
<th>Q-index</th>
<th>Scalability</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.17</td>
<td>0.97</td>
<td>-1.64</td>
<td>0.25</td>
<td>1.03</td>
<td>0.99</td>
<td>0.26</td>
<td>78.43</td>
<td>-1.98</td>
</tr>
<tr>
<td>2</td>
<td>0.18</td>
<td>1.42</td>
<td>-0.43</td>
<td>0.24</td>
<td>0.90</td>
<td>0.96</td>
<td>0.21</td>
<td>108.53</td>
<td>-1.05</td>
</tr>
<tr>
<td>3</td>
<td>0.24</td>
<td>1.18</td>
<td>-2.45</td>
<td>0.00</td>
<td>0.95</td>
<td>0.92</td>
<td>0.22</td>
<td>78.81</td>
<td>-2.69</td>
</tr>
<tr>
<td>4</td>
<td>0.21</td>
<td>1.84</td>
<td>-2.63</td>
<td>0.00</td>
<td>0.66</td>
<td>0.88</td>
<td>0.15</td>
<td>73.93</td>
<td>-3.72</td>
</tr>
<tr>
<td>5</td>
<td>0.56</td>
<td>1.85</td>
<td>-2.31</td>
<td>0.00</td>
<td>0.64</td>
<td>0.86</td>
<td>0.13</td>
<td>90.28</td>
<td>-3.26</td>
</tr>
<tr>
<td>6</td>
<td>0.23</td>
<td>1.64</td>
<td>-2.62</td>
<td>0.00</td>
<td>0.58</td>
<td>0.89</td>
<td>0.16</td>
<td>76.58</td>
<td>-3.50</td>
</tr>
<tr>
<td>7</td>
<td>0.36</td>
<td>0.69</td>
<td>-3.54</td>
<td>0.00</td>
<td>1.30</td>
<td>1.01</td>
<td>0.30</td>
<td>51.72</td>
<td>-2.59</td>
</tr>
<tr>
<td>8</td>
<td>0.26</td>
<td>1.54</td>
<td>-1.56</td>
<td>0.11</td>
<td>0.72</td>
<td>0.88</td>
<td>0.17</td>
<td>108.38</td>
<td>-2.17</td>
</tr>
<tr>
<td>9</td>
<td>0.19</td>
<td>1.90</td>
<td>0.12</td>
<td>0.19</td>
<td>0.88</td>
<td>0.91</td>
<td>0.19</td>
<td>126.58</td>
<td>-0.33</td>
</tr>
<tr>
<td>10</td>
<td>0.52</td>
<td>1.06</td>
<td>-2.66</td>
<td>0.00</td>
<td>0.91</td>
<td>0.94</td>
<td>0.23</td>
<td>74.00</td>
<td>-2.71</td>
</tr>
<tr>
<td>11</td>
<td>0.29</td>
<td>1.60</td>
<td>-2.82</td>
<td>0.00</td>
<td>0.72</td>
<td>0.89</td>
<td>0.17</td>
<td>68.77</td>
<td>-3.71</td>
</tr>
<tr>
<td>12</td>
<td>0.34</td>
<td>0.28</td>
<td>-2.20</td>
<td>0.00</td>
<td>1.45</td>
<td>1.24</td>
<td>0.39</td>
<td>34.14</td>
<td>-0.74</td>
</tr>
<tr>
<td>13</td>
<td>0.38</td>
<td>1.55</td>
<td>-1.97</td>
<td>0.00</td>
<td>0.69</td>
<td>0.87</td>
<td>0.16</td>
<td>102.28</td>
<td>-2.54</td>
</tr>
<tr>
<td>14</td>
<td>0.26</td>
<td>2.11</td>
<td>-0.71</td>
<td>0.16</td>
<td>0.72</td>
<td>0.84</td>
<td>0.14</td>
<td>137.78</td>
<td>-1.36</td>
</tr>
<tr>
<td>15</td>
<td>0.46</td>
<td>1.06</td>
<td>-1.60</td>
<td>0.00</td>
<td>1.06</td>
<td>0.94</td>
<td>0.23</td>
<td>96.23</td>
<td>-1.65</td>
</tr>
<tr>
<td>16</td>
<td>0.32</td>
<td>1.49</td>
<td>-1.57</td>
<td>0.06</td>
<td>0.76</td>
<td>0.87</td>
<td>0.17</td>
<td>110.66</td>
<td>-2.07</td>
</tr>
<tr>
<td>17</td>
<td>0.35</td>
<td>0.48</td>
<td>-1.76</td>
<td>0.00</td>
<td>1.31</td>
<td>1.15</td>
<td>0.34</td>
<td>55.32</td>
<td>-0.97</td>
</tr>
<tr>
<td>18</td>
<td>0.33</td>
<td>1.85</td>
<td>-1.95</td>
<td>0.00</td>
<td>0.62</td>
<td>0.85</td>
<td>0.14</td>
<td>105.02</td>
<td>-2.73</td>
</tr>
<tr>
<td>19</td>
<td>0.26</td>
<td>1.39</td>
<td>-0.50</td>
<td>0.34</td>
<td>0.93</td>
<td>0.98</td>
<td>0.23</td>
<td>98.76</td>
<td>-1.33</td>
</tr>
<tr>
<td>20</td>
<td>0.32</td>
<td>1.90</td>
<td>-0.05</td>
<td>0.18</td>
<td>0.84</td>
<td>0.89</td>
<td>0.17</td>
<td>131.20</td>
<td>-0.51</td>
</tr>
</tbody>
</table>
Table 5.1: Item indicators for Duval data (continued)

<table>
<thead>
<tr>
<th>Item</th>
<th>% Checks</th>
<th>3pl Disc</th>
<th>3pl Diff</th>
<th>3pl Guess</th>
<th>OMS</th>
<th>IMS</th>
<th>Q-index</th>
<th>Scalability</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>0.24</td>
<td>1.44</td>
<td>0.02</td>
<td>0.20</td>
<td>0.95</td>
<td>0.97</td>
<td>0.22</td>
<td>111.37</td>
<td>-0.48</td>
</tr>
<tr>
<td>22</td>
<td>0.45</td>
<td>1.94</td>
<td>-0.32</td>
<td>0.23</td>
<td>0.81</td>
<td>0.89</td>
<td>0.17</td>
<td>128.57</td>
<td>-0.96</td>
</tr>
<tr>
<td>23</td>
<td>0.43</td>
<td>2.03</td>
<td>-0.35</td>
<td>0.50</td>
<td>0.84</td>
<td>0.97</td>
<td>0.22</td>
<td>97.86</td>
<td>-1.69</td>
</tr>
<tr>
<td>24</td>
<td>0.35</td>
<td>2.12</td>
<td>-0.93</td>
<td>0.21</td>
<td>0.67</td>
<td>0.84</td>
<td>0.14</td>
<td>130.36</td>
<td>-1.75</td>
</tr>
<tr>
<td>25</td>
<td>0.36</td>
<td>1.85</td>
<td>-0.97</td>
<td>0.25</td>
<td>0.74</td>
<td>0.87</td>
<td>0.16</td>
<td>120.34</td>
<td>-1.79</td>
</tr>
<tr>
<td>26</td>
<td>0.30</td>
<td>2.08</td>
<td>0.46</td>
<td>0.28</td>
<td>0.99</td>
<td>0.99</td>
<td>0.23</td>
<td>104.98</td>
<td>-0.22</td>
</tr>
<tr>
<td>27</td>
<td>0.36</td>
<td>0.87</td>
<td>0.34</td>
<td>0.22</td>
<td>1.14</td>
<td>1.11</td>
<td>0.30</td>
<td>74.73</td>
<td>-0.30</td>
</tr>
<tr>
<td>28</td>
<td>0.23</td>
<td>2.25</td>
<td>0.15</td>
<td>0.14</td>
<td>0.83</td>
<td>0.86</td>
<td>0.16</td>
<td>139.47</td>
<td>-0.18</td>
</tr>
<tr>
<td>29</td>
<td>0.26</td>
<td>1.82</td>
<td>-0.77</td>
<td>0.20</td>
<td>0.76</td>
<td>0.87</td>
<td>0.16</td>
<td>127.32</td>
<td>-1.45</td>
</tr>
<tr>
<td>30</td>
<td>0.30</td>
<td>0.94</td>
<td>0.91</td>
<td>0.14</td>
<td>1.13</td>
<td>1.09</td>
<td>0.29</td>
<td>78.30</td>
<td>0.38</td>
</tr>
<tr>
<td>31</td>
<td>0.32</td>
<td>1.46</td>
<td>1.06</td>
<td>0.22</td>
<td>1.16</td>
<td>1.08</td>
<td>0.29</td>
<td>77.96</td>
<td>0.42</td>
</tr>
<tr>
<td>32</td>
<td>0.34</td>
<td>1.31</td>
<td>2.29</td>
<td>0.31</td>
<td>1.42</td>
<td>1.26</td>
<td>0.39</td>
<td>29.43</td>
<td>0.67</td>
</tr>
<tr>
<td>33</td>
<td>0.27</td>
<td>1.73</td>
<td>-0.84</td>
<td>0.30</td>
<td>0.80</td>
<td>0.90</td>
<td>0.18</td>
<td>114.21</td>
<td>-1.69</td>
</tr>
<tr>
<td>34</td>
<td>0.27</td>
<td>1.37</td>
<td>-0.04</td>
<td>0.25</td>
<td>0.98</td>
<td>0.99</td>
<td>0.23</td>
<td>103.78</td>
<td>-0.68</td>
</tr>
<tr>
<td>35</td>
<td>0.34</td>
<td>0.47</td>
<td>0.63</td>
<td>0.01</td>
<td>1.23</td>
<td>1.18</td>
<td>0.33</td>
<td>57.67</td>
<td>0.32</td>
</tr>
<tr>
<td>36</td>
<td>0.18</td>
<td>1.57</td>
<td>0.04</td>
<td>0.15</td>
<td>0.92</td>
<td>0.93</td>
<td>0.20</td>
<td>122.25</td>
<td>-0.33</td>
</tr>
<tr>
<td>37</td>
<td>0.34</td>
<td>1.25</td>
<td>1.34</td>
<td>0.22</td>
<td>1.21</td>
<td>1.11</td>
<td>0.31</td>
<td>67.34</td>
<td>0.57</td>
</tr>
<tr>
<td>38</td>
<td>0.28</td>
<td>2.65</td>
<td>0.07</td>
<td>0.21</td>
<td>0.81</td>
<td>0.86</td>
<td>0.16</td>
<td>140.17</td>
<td>-0.46</td>
</tr>
<tr>
<td>39</td>
<td>0.20</td>
<td>2.19</td>
<td>0.04</td>
<td>0.22</td>
<td>0.84</td>
<td>0.89</td>
<td>0.17</td>
<td>132.32</td>
<td>-0.51</td>
</tr>
<tr>
<td>40</td>
<td>0.24</td>
<td>1.22</td>
<td>2.05</td>
<td>0.12</td>
<td>1.40</td>
<td>1.09</td>
<td>0.33</td>
<td>52.15</td>
<td>1.49</td>
</tr>
</tbody>
</table>
Table 5.1: Item indicators for Duval data (continued)

<table>
<thead>
<tr>
<th>Item</th>
<th>% Checks</th>
<th>3pl Disc</th>
<th>3pl Diff</th>
<th>3pl Guess</th>
<th>OMS</th>
<th>IMS</th>
<th>Q-index</th>
<th>Scalability</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>0.26</td>
<td>1.18</td>
<td>0.45</td>
<td>0.21</td>
<td>1.07</td>
<td>1.05</td>
<td>0.26</td>
<td>91.60</td>
<td>-0.14</td>
</tr>
<tr>
<td>42</td>
<td>0.19</td>
<td>2.03</td>
<td>1.40</td>
<td>0.44</td>
<td>1.28</td>
<td>1.22</td>
<td>0.35</td>
<td>46.31</td>
<td>-0.11</td>
</tr>
<tr>
<td>43</td>
<td>0.42</td>
<td>1.24</td>
<td>-0.05</td>
<td>0.18</td>
<td>0.98</td>
<td>0.99</td>
<td>0.23</td>
<td>106.90</td>
<td>-0.50</td>
</tr>
<tr>
<td>44</td>
<td>0.29</td>
<td>1.55</td>
<td>0.99</td>
<td>0.13</td>
<td>1.06</td>
<td>0.99</td>
<td>0.24</td>
<td>95.32</td>
<td>0.73</td>
</tr>
<tr>
<td>45</td>
<td>0.15</td>
<td>3.41</td>
<td>-0.28</td>
<td>0.24</td>
<td>0.69</td>
<td>0.80</td>
<td>0.12</td>
<td>152.05</td>
<td>-1.04</td>
</tr>
<tr>
<td>46</td>
<td>0.55</td>
<td>1.40</td>
<td>0.79</td>
<td>0.19</td>
<td>1.06</td>
<td>1.03</td>
<td>0.25</td>
<td>92.82</td>
<td>0.27</td>
</tr>
<tr>
<td>47</td>
<td>0.15</td>
<td>2.05</td>
<td>0.80</td>
<td>0.16</td>
<td>1.01</td>
<td>0.96</td>
<td>0.22</td>
<td>105.91</td>
<td>0.53</td>
</tr>
<tr>
<td>48</td>
<td>0.15</td>
<td>2.61</td>
<td>1.03</td>
<td>0.31</td>
<td>1.17</td>
<td>1.09</td>
<td>0.29</td>
<td>74.33</td>
<td>0.27</td>
</tr>
<tr>
<td>49</td>
<td>0.13</td>
<td>1.39</td>
<td>1.14</td>
<td>0.29</td>
<td>1.18</td>
<td>1.12</td>
<td>0.31</td>
<td>69.47</td>
<td>0.22</td>
</tr>
<tr>
<td>50</td>
<td>0.15</td>
<td>1.89</td>
<td>0.12</td>
<td>0.33</td>
<td>0.92</td>
<td>0.97</td>
<td>0.21</td>
<td>111.43</td>
<td>-0.72</td>
</tr>
<tr>
<td>51</td>
<td>0.16</td>
<td>2.29</td>
<td>0.93</td>
<td>0.22</td>
<td>1.07</td>
<td>1.00</td>
<td>0.25</td>
<td>93.86</td>
<td>0.50</td>
</tr>
<tr>
<td>52</td>
<td>0.16</td>
<td>1.74</td>
<td>1.02</td>
<td>0.27</td>
<td>1.14</td>
<td>1.08</td>
<td>0.28</td>
<td>78.40</td>
<td>0.26</td>
</tr>
<tr>
<td>53</td>
<td>0.46</td>
<td>1.15</td>
<td>1.08</td>
<td>0.22</td>
<td>1.17</td>
<td>1.11</td>
<td>0.30</td>
<td>72.92</td>
<td>0.34</td>
</tr>
<tr>
<td>54</td>
<td>0.24</td>
<td>1.40</td>
<td>0.67</td>
<td>0.13</td>
<td>1.02</td>
<td>0.99</td>
<td>0.23</td>
<td>102.90</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Figure 5.2: Fit Statistic Histograms: Solid black curve gives density curve of fit statistics for Duval data. Solid vertical lines show mean and plus/minus one standard deviation from the mean for this empirical density. The dotted lines are the mean and plus/minus one standard deviation for fit statistics based on Rasch data simulated to be similar to the Duval data. The fact that nearly all of the empirical distribution is outside of this range suggests that the Rasch model is not a plausible data generating model.
data is clearly not conforming to the assumptions of the Rasch model.

The small multiples plot for the Lexile data is shown in Figure 5.3. All of the fit statistics were standardized for the Figure. The two black lines are at plus and minus one standard deviations for the distribution of fit statistics. As a consequence, deviant items can only be determined in a relative sense (internal to each test). So items that may be problematic in an absolute sense but that are less deviant than the other items on on this instrument wouldn’t be noticeable. Looking through the items, there are some parallels to what was seen before with the simulated data. Item 7, for example, clearly has a problem with guessing. Interestingly, there are few if any items that present the “shoulder” pattern (higher bumps at high and low abilities) associated with non-unity discriminations as were observed in simulation. In contrast, we see a number of items that show slow upward growth in the proportion of distortions (observed item 10 or 38, for example). This could point to a break down of the scale at higher abilities since cells in this part of the scale are not consistent with the measurement axioms at especially high rates. This is an intriguing result.

I tried to compare the fit statistics and the checks in other ways, for example by looking at which items would be flagged using both indices, but since all items had at least 10% of their checks showing violations (and this was 5 times worse than the 2% thresholds set with true Rasch data) there was little to say. Rather than focus on this comparison, it is perhaps more interesting to ask whether specific item-side deviations may be responsible for the data not conforming to the Rasch model.

**Fitting the 3pl** To determine whether item-side deviations are driving the fit problems indicated above, I fit the 3pl to the Duval data. After fitting the 3pl instead of the Rasch model, one observes impressive variability in terms of both the discriminations and guessing parameters, see Figure 5.4. The discrimination parameters range from 0.3 to 3.4. Twenty six items have guessing parameters above 0.2, the guessing parameter that we might expect if there are 3 or 4 distractors for each cloze item. The 3pl difficulty estimates are over-estimates of the Rasch difficulty estimates, see the middle panel of Figure 5.4. This over-estimation is due to the fact that without the guessing parameters, correctly guessed responses are being interpreted as signs of ability in the
Figure 5.3: Small multiples for Duval data
Rasch model. Hence, lower ability students look more able and the item looks easier. Recall the variability in the CSAP item parameters that I discussed last chapter in the treatment simulation. The variability in the estimated 3pl item parameters from the Duval data is worse than on tests that specifically use the 3pl! Clearly assuming that the guessing parameters are all identically zero and the discrimination parameters all identically unity was far too optimistic a set of assumptions for the Duval data.

Figure 5.5 compares the proportion of violations (the y-axis) to the estimated item parameters (the x-axis). There are two things one might expect to see. First, one might expect discrimination parameters farther from unity to be associated with more detected violations. Second, one might expect guessing parameters farther from zero to be associated with more detected violations. Neither of these seems to be the case, although perhaps there are fewer violations being detected for the larger item discriminations. Consequently, it seems that while these item-side deviations are presumably responsible for some of the lack of fit with the Rasch model there may well be other sources of the poor fit. Taken as a whole, the evidence suggests that the data here is not even close to being generated by the Rasch model. Treating the resulting ability scale as interval would certainly not be justified by either ACM or IRT.

5.2 State Data

I now turn to data generated from an anonymous state’s NCLB mandated assessment for Reading. States typically use either the Rasch (1960) model or Birnbaum’s 3pl (1968) model for the purposes of scaling such assessments (this choice varies partially as a function of test developer). This state uses the Rasch model, which is advantageous since it then necessitates only the more straightforward ACM/Rasch analogy rather than the introduction of polynomial conjoint measurement (see Kyngdon, 2011).

There are two additional unique features of this test and data. The first is that the test is computer-adaptive. Although this has a number of ramifications, from the perspective of the checks the most interesting one is that students do not take the same number of items. This complicates
Figure 5.4: Item Parameters: Analysis of 3pl parameters estimated from supposedly Rasch Duval data.
Figure 5.5: Item Parameters as compared to the proportion of failed checks.
matters since sum scores are going to be used for grouping students. The second unique aspect is that this test is vertically scaled via cross-grade common items. Consequently, in the data there are responses from students in different grades to the same item (this is known as a non-equivalent group common item design, see Kolen & Brennan, 2004). The Lexile scale is also a vertical scale but the Duval, NC data used here consisted of responses from students in only a single grade. Inclusion of children in pairs of grades allows for focus on the so-called common items that are used to create the vertical scale. If these items are consistent with the axioms, then this will give some indication that the vertical scale (as opposed to merely the separate scales created for each grade) is perhaps interval.

5.2.1 Data

The adaptive data is challenging to work with for my purposes due to the structural miss- ingness. Students need to respond to a common number of items if their sum scores are to be used for ordering. With this data, however, students not only respond to different numbers of items, but the items they respond to also vary. One potential approach would be to use ability estimates generated from all responses and then use those to order abilities in “blocks” of items. If single cancellation holds (and individuals can be ordered without reference to the specific items), this is perfectly reasonable. The problem comes in that students need to be grouped into common ability levels. With Rasch ability estimates from traditional assessment data (all students respond to the same items), there is no problem as there is a 1–1 correspondence between sum scores and ability estimates. Here there is no such guarantee. Decisions have to be made about how to group students based on the abilities and this invites a level of subjectivity into the analysis. For this reason, I focus on ability estimates via sum scores from small blocks of items that contain full sets of responses for groups of students. Though these blocks contain fewer items than the analyses I’ve done up till this point, they allow for internally coherent analyses.

The dataset I use consists of only common items. These are of crucial importance in evaluating the quality of the resultant scale since if these items do not support an interval interpretation
then it is impossible to see how the vertical scale will. There are differing numbers of common items between pairs of grades. My strategy is to take as many common items as possible between a pair of grades that contain complete responses sequences from a reasonably large number of students (at least 1,000 students taken from both grades). These restrictions, combined with the availability of item difficulty estimates\(^8\), restricted attention to the grade pairs 4–5 and 5–6. For each pair, I focused on 8 common items.\(^9\) For the 4–5 data, there were 1,138 students (574 from grade 4 and 564 from grade 5) with usable information. For the 5–6 pair, there were 1,098 students (561 from grade 5 and 537 from grade 6). I re-emphasize that these blocks contain full response strings (no missing data), which is important since missing data makes grouping respondents a challenge.

### 5.2.2 Results

Results for grade 4–5 and 5–6 are shown in Figure 5.6. Compared to what we observed for the Lexile data, these results are clearly far more supportive of the cancellations axioms of conjoint measurement. Moreover, the spikes on the left hand side for a sum score of 1 are not too worrisome since these sum scores had very few respondents. Even with those spikes, the overall unweighted means for the checks in each conjoint matrix were 0.09 and 0.06. Weighted means were lower still, 0.05 and 0.02. The latter is not far from what was observed in the case of simulated Rasch data. Weighted means for each item are shown on the right hand side of each panel in Figure 5.6.

One possible hypothesis about why the data is seeming to satisfy the axioms so much better than with the Lexile data is that there are only 8 items in each case. However, I do not think the operationalized versions of the checks used here is sensitive to the number of items in question since the procedure revolves around the sampling of a 3-matrix from a data matrix. With a 3-matrix, one is only handling 3 items at 3 sum scores. In contrast, it’s possible that fit statistics computed at the item level should be sensitive to the number of items since they are making a more global

---

\(^8\) These estimates were computed on the basis of the full adaptive assessment. They are usable, as opposed to the individual ability estimates, since there is no need to group items.

\(^9\) With only 8 items, there are at most \(\binom{8}{3} \binom{7}{3} = 1960\) 3-matrices of interest (only 7 sum scores since neither 0 or 8 would be interesting). Rather than sample each 3-matrix with certainty, a possibility in this case due to the reasonable number of matrices, I still rely on the sampling approach primarily so that I can utilize my already developed software.
Figure 5.6: Results of Checks for state data
statement about whether curves are crossing anywhere on the ability spectrum.

This point about curves leads to the following. The checks are determining whether there are regions in which empirical item response functions (or person response functions) are crossing. Rasch fit statistics are meant to examine a stronger question: whether the Rasch model is the data generating model. To answer this strong question, a complete battery of fit statistics would be attempting to determine whether the empirical response function is consistent with a certain parametric form (logistic?, constant discrimination, lower asymptote at zero, monotonically increasing) of that response function. I make this point for the following reason. The checks seem to suggest that an interval scaling of the state data may be possible. However, the fit statistics may suggest that the Rasch model does not fit. In that case, I would claim that the interval scale known to exist as a consequence of Luce & Tukey, 1964 is not the one produced by the Rasch model.

A traditional approach to interpreting fit statistics can be performed using Figure 5.7. This Figure compares the proportion of sampled 3-matrices for each item (in the 5–6 group, chosen since it seems to be the more promising candidate for interval scaling based on the checks) that showed distortions to the fit statistics. I focus on infit, outfit, and the Q-index ("i", "o", and "q") and ignoring the scalability coefficient due to its weak performance in the simulation from the prior Chapter. The especially poor behavior on the left hand side in several of the figures is due to very small sample sizes (looking back at Figure 5.2, these contribute almost nothing to the weighted means). Using values of 0.8 and 1.2 for the mean-square statistics, items 2, 3, 6, and 7 would be flagged due to infit values below 0.8. The outfit statistics are all acceptable. These thresholds are too lax, however, given the sample size. Given the samples of roughly 1000 students, 0.9 and 1.1 would be more appropriate (see Table 1 of Wu & Adams, n.d.). Using those critical values, every item would be flagged for either a deviant infit or outfit statistic. There are no ready guidelines for using the fit statistic of Rost and von Davier (1994), but numbers near 0 indicate perfect discrimination while numbers near 0.5 indicate that the item and latent trait of interest are independent. The fit statistics shown here are fairly low (lower, for example, than those shown in the example in Table 3 of Rost & von Davier, 1994) and I don’t think they would necessitate any
item being flagged.

Any handling of fit statistics that relies upon external criteria will be suboptimal. For example, suppose simulations have shown that certain critical values apply when working with fit statistics computed for assessments with 10 items administered to 1000 students. An instrument that happens to have the same number of students and items may perhaps require slightly different thresholds due to a different concentration of item difficulties. Since the state data is relatively small, I choose to also simulate fit statistic distributions. This method of assessing the fit of the Rasch model is not new. R. Smith (2010) advocated such an approach and has relied upon it in his own work (R. M. Smith, 1991). Not only does it have a history in the psychometrics literature, but it also bears some resemblance to Gelman’s ideas on model checking via simulation (e.g., Gelman et al., 2004). The key element in each is that the stochastic model is taken quite seriously. Since it is being posed as the generative mechanism for reality, then it should be able to reproduce reality. If one can not replicate reality to a desired approximation, then this is treated as evidence that the model is flawed. The nature of the flaw may be discernible in observing how the simulated data departs from the real data.

I took the following approach to simulate the distribution of fit statistics. First, using the estimated abilities and item parameters from the common item data for the 5–6 pair of grades, I simulated 1000 sets of responses. For each of these, estimated abilities and difficulties were then generated. Using these estimates, I finally computed fit statistics for each of the 1000 iterations. The fit statistics from the 1000 iterations are then used to form a distribution. The question is then where the empirically observed fit statistics fall in this simulated distribution. If, for example, an empirically observed fit statistic falls at the median, then this is evidence that the stochastic model generates data that “looks” (in terms of the fit statistic at least) like what was actually observed. If the empirically observed fit statistic is at the outer reaches of the distribution, this is evidence that the data generated by the stochastic model and the empirical data are not congruent.

Table 5.2 shows the estimated quantile of each item fit statistic produced from the empirical cumulative distribution function created over the 1000 iterations of the simulation process described
Figure 5.7: Small Multiple for state showing checks versus fit statistics
Table 5.2: Fit statistics for state: Quantile of empirical fit statistic relative to a distribution of simulated fit statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>Outfit</th>
<th>Infit</th>
<th>Q-index</th>
<th>Scalability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.71</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
</tr>
<tr>
<td>3</td>
<td>0.99</td>
<td>0.78</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.82</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.99</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0.78</td>
<td>0.14</td>
<td>0.92</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0.38</td>
<td>0.05</td>
<td>0.84</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
above. Values of 1 or 0 indicate that the empirical fit statistic was either larger or smaller than all of the fit statistics estimated from empirical data. Based on the number of extreme fit statistics of this type (or values near 0 or 1), I see no other choice but to conclude that the Rasch model did not generate the empirical data. As an aside, it seems as though in fit is the least stringent measure (the measure least capable of detecting differences between the model and the real empirical data), followed by outfit and the Q-index, and the scalability coefficient is the most stringent. A few of the items seem nearly consistent with the Rasch model in terms of these fit statistics. Items 6 & 7 are promising, especially if one doesn’t mind ignoring the scalability coefficient. Considering the disappointing results I uncovered for that statistic in the previous chapter, it may be reasonable to ignore it. In summary, the state data seems to largely obey the cancellation axioms but does not conform to the more rigorous demands of the Rasch model.

5.3 Conclusions

In this Chapter, I investigated whether two empirical datasets that have been scaled with the Rasch model actually meet the cancellation axioms as tested by the checks. The Duval data fell woefully short of meeting the cancellation axioms. The state data seemed to meet the axioms but was not generated by the Rasch model. Hence, these are empirical examples of the difference between cells II and III in Table 5.3. The state data, in cell II, is such that a quantitative variable may reasonably be conjectured as driving the responses. The Duval data with the Lexile scale, in cell III, does not support such a hypothesis. Knowing the quality of data in this sense, perhaps combined with what is known from an analysis of fit statistics, could be quite useful. For example: since the Rasch model does not fit the state data well, it would be prudent to investigate whether there were more flexible models that could be used to fit the data. If guessing was thought to be the problem, perhaps a single guessing parameter could be used for all items. This modification would not induce crossing curves since all the items still have a common lower asymptote and interval scales for this model could still be motivated using the simple version of ACM presented here. Additional argumentation may be required to justify interval interpretations if the model modifications are
Table 5.3: Comparison of Decisions based on Rasch fit statistics and the checks

<table>
<thead>
<tr>
<th>Rasch Fit Statistics</th>
<th>RM fits</th>
<th>RM does not fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Checks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>consistent with ACM</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>not consistent with ACM</td>
<td>IV</td>
<td>III</td>
</tr>
</tbody>
</table>

more complex. The point is that the checks confirm that an interval interpretation is possible. With the Lexile data, things are bleaker. The data itself needs to be refined.

One probable reason for the poor performance of the Duval data could be the reliance of their theory on reading comprehension to construct instruments without any kind of check on item performance. When items are calibrated based solely on the theorized comprehension difficulty of the stem, this does not necessarily guarantee a well functioning item. It would be instructive to compare the results from this analysis to those obtained from data where items were constructed in the standard Lexile approach but then pre-screened for consistency with the Rasch model (via fit statistics). If data produced under such an approach yielded more favorable results in terms of the checks, this would make for a viable, and seemingly superior, alternative.

With respect to the state data, my analysis focused on a smaller set of items than in the earlier simulation work. This was done partially due to the difficulties created by the adaptive design but also to focus on the cross-grade scale linkage. On the one hand, this restriction is such that I can not state conclusively that the scale formed from all the items is interval since I have only analyzed a relatively small subset. On the other hand, since this subset was composed of students from multiple grades this is tantalizing evidence that the vertical scale (not just the grade-specific subscales) is potentially interval. This latter result is preliminary but well worth the required restriction.
If interval scales are possible in educational assessment, this would make the analysis of scale scores and communication about scale scores much easier. The most commonly used method for scale construction in educational assessment is item response theory (IRT). As even the founders of IRT noted (e.g., Lord & Novick, 1968, p. 20–21), these methods do not necessarily produce interval scales. A lesser known measurement paradigm is additive conjoint measurement (ACM). ACM provides a means of arguing that interval scales are possible for a pair of latent variables that underlie certain data. This approach is non-constructive, however, in that it does not provide a method for converting response data to scale scores. This dissertation examines a particular technique for analyzing the veracity of the axioms of ACM in empirical educational assessment data.

It is quite challenging to integrate the machinery of conjoint measurement with assessment data since the latter contains measurement error. A workable approach was suggested by Karabatsos (2001), although in the course of this dissertation it has been shown that this version of the methodology was flawed in several respects. The first problem was that the methodology wrongly used the middle cell of a 3-matrix to impose ordering along the minor diagonal. The second problem was that the ordering along the minor diagonal (remember that the conclusion of the double cancellation conditional requires a certain ordering of the corner cells) was imposed without any attempt to make sure that the premise of the double cancellation conditional was true. I have described a modified version of the methodology that corrects these two issues. Although this was
not one of the original research goals, the methodological improvement is an important contribution of this dissertation. The methodological improvement was paired with a qualitatively different approach to the checks than either Karabatsos (2001) or Kyngdon (2011) in that I sampled vastly greater numbers of 3-matrices. Where they analyze one or two 3-matrices, I analyzed thousands. Conceptually, this is important as not only am I analyzing a much greater portion of the conjoint matrix but it is also important to analyze cells in different configurations since it is only in certain configurations that violations can be exposed. This different approach led to an important discovery. Compare the results of Kyngdon’s (2011) study of the Lexile scale to the results from Chapter 5: where he found no violations and preliminarily concluded that reading ability was a quantity, my analysis found numerous violations. Based on these violations, I argued that it is premature to conclude that reading ability as conceptualized by the Lexile theory is a quantity.

On top of the methodological refinements, there were several specific research questions that I attempted to answer. The first research question was whether the checks were an effective statistical tool ([Q1] from Chapter 1), an issue that had been overlooked by Karabatsos (2001) and Kyngdon (2011). I worked with simulated data so that it would be known whether the generating model was Rasch or not. The point being that in the case of Rasch data, interval scales are possible. I was able to demonstrate that in the case of purely Rasch data, the checks detected violations in roughly 2% of the checked cells. With non-Rasch data, this mean percentage of violations was higher. Guessing or discrimination parameters that deviated from the Rasch model could frequently be observed by looking at the percentage of violations as a function of ability.

One potentially discouraging finding was that the checks were not able to discriminate between items with and without deviant guessing parameters when one focused solely on the mean (consider Table 4.1 for an illustration of this fact). The problem seemed less serious when one looked at the pattern of violations as a function of ability, in that case violations could be detected in at least some cases. But this is a time consuming task. This finding is particularly concerning since guessing seemed to be driving the most serious scale distortions observed via the difference

---

1 At the time of my prospectus defense, I had not yet become aware of these problems.
matrices. Consequently, the checks are less able to detect the violations that cause the largest distortions. In practice, the danger of missing major scale distortions can be mitigated if data is examined using the cancellation axioms in tandem with traditional fit statistics. Suppose that the checks indicate data is consistent with the cancellation axioms. Then the next stage of fitting the Rasch model should be performed with extra scrutiny (via fit statistics and empirical item response curves) on whether guessing seems to be a problem.

A subquestion of the first research question was what, if anything, the checks added to our knowledge of misfit when they are combined with traditional Rasch fit statistics [Q1A]. The most useful information that one gets from considering the checks is from graphics such as Figures 4.2 and 4.3. In these Figures, one can identify the type of deviation fairly rapidly. If all the violations are at the low end, guessing is probably a problem. When the violations occur on both sides but not in the middle, discrimination is probably an issue. Such patterns were observed with empirical data as well, although patterns that did not correspond to either item-side deviation also occurred. In Figure 5.3 for example, the proportion of observed violations sometimes rose along with the ability level. This is a rather odd pattern. From a modeling perspective, the only fix I can see would be a missing upper asymptote for the probability of a correct response along the lines suggested in Barton and Lord (1981). This is a rarely used modification to the item response models in educational assessment and, in the case of the Lexile data, would certainly not be the first problem one would want to address. It would be interesting to know if this pattern was present in other empirical examples.

A second research question was whether specific deviations from the assumptions of certain IRT models resulted in meaningful scale distortions [Q2]. My initial exploration of this question involved the difference matrices. Using these, I was able to demonstrate that estimated abilities from the low end of the scale would see more compression than estimates from the high end of the scale relative to the true abilities when estimation occurred in the presence of certain item-side deviations. This means that a difference of \( \delta \) in the estimated scores at the low end corresponds to a larger difference in true abilities than a difference of \( \delta \) at the high end of the estimated ability scale.
This is meaningful since it demonstrates that low and high abilities are differentially impacted. However, it was challenging to leverage this result at the score-level into a policy-relevant finding. The treatment and clustering simulations were both attempts to do so. Each of these simulations demonstrated the same pattern. For the treatment simulation, the estimated treatment effect was lower (as judged by statistical significance in many cases) for lower ability students even though the true effect was constant. Though the interaction was frequently significant, its magnitude was relatively small.

The clustering simulation introduced a method of clustering students into schools using a multinomial distribution. This method was able to replicate the kinds of between- versus within-school variability in student abilities that are actually observed in the state of Colorado. Due to the systematic effect noted above, I was able to demonstrate that the variability in student scores was under-estimated relative to truth for schools with lower ability on the mean while the variability in higher ability schools was over-estimated. The differences between true and estimated quantities were again rather small. Neither of these two simulations suggested that scale distortions would be meaningful in the associated policy contexts.

Although these analyses suggested that the scale distortions weren’t likely to be huge, the analyses were not exhaustive. Both simulations contextualized policy-relevance as being based on the aggregation of individual level scores (whether it be into control and treatment groups or into schools). Such aggregation is quite common in educational research and frequently plays a role in education policy decision-making. However, differences due to scale distortions at the level of the individual, not the primary focus here, could be meaningful in some cases. This is a potentially serious issue that merits further investigation.

If test creators use the Rasch model for scaling, they clearly believe their data fits the Rasch model. Therefore, it should also meet the axioms of conjoint measurement. The third research question actually put this assumption to the test: are certain datasets scaled with the Rasch model consistent with the axioms of conjoint measurement [Q3]? Once the methodology had been validated as behaving appropriately with simulated data, it was appropriate to check two empirical
data sets. The first set of data is based on the Lexile scale (Stenner et al., 1983). Perhaps the most interesting feature of this scale is that difficulties can be estimated before the items are administered via a theory of reading difficulty developed by MetaMetrics. This scale was developed to be interval. The goal is to make a scale to measure reading difficulty that is identical to a thermometer for measuring temperature. In fact, Stenner (1996) claimed that the Lexile scale achieved such an objective. Kyngdon (2011) offered qualified support for this proclamation as well based on using the Karabatsos (2001) version of the methodology.

My analysis of the data from Kyngdon (2011) suggests that Lexile scale scores do not yet merit being treated as interval. The checks detected violations at 10 times the rate that detections were detected in truly Rasch data. Not surprisingly, fit statistics also suggested that the data was not consistent with the Rasch model. The primary problem seemed to be that the theory of reading difficulty, though it may be predictive of item difficulty, did nothing to ensure that items have identical discrimination or guessing parameters. Indeed, why would the theory have ensured such behavior from the items? It would make sense for the “cloze” items that make up the Lexile scale to be pre-screened for these undesirable item characteristics before being administered as part of an operational test if interval scales are truly desired.

The second empirical data set was for an anonymous state. This data was challenging to work with since it came from a computer-adaptive administration and also contained responses from students in different grades. The latter feature was especially interesting since it allowed for a focus on the so-called “common items”, items embedded in multiple grades that are used to place the scale scores from different test forms on a common scale. I analyzed two grade pairs and found that both were producing violations of the measurement axioms at much lower rates than the Lexile data, possibly due to the fact that misfitting items had already been pruned, with one of the pairs producing violations at about the level one would expect from truly Rasch data. However, an analysis of fit statistics suggested that although this data may be consistent with the measurement axioms it did now show adequate fit with the Rasch model.

This finding is interesting. To make sense of it, consider again the framework of Table 6.1.
Table 6.1: Comparison of Decisions based on Rasch fit statistics and the checks

<table>
<thead>
<tr>
<th>Rasch Fit Statistics</th>
<th>RM fits</th>
<th>RM does not fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Checks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>consistent with ACM</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>not consistent with ACM</td>
<td>IV</td>
<td>III</td>
</tr>
</tbody>
</table>

Recall that the first task is to first find evidence that the data is consistent with the measurement axioms. If so, then ACM implies that interval scales can be created for person ability and item difficulty. If such evidence can be found, then data is in at least cell II. The second task is then to examine the fit of the Rasch model to the data. A subset of data that meet the axioms of conjoint measurement will also meet the assumptions of the Rasch model. Interval scales for this subset of data, in cell I, can be created via the Rasch model. Data in cell II leads to a more challenging scenario since ACM implies that interval scales are possible and yet the Rasch model’s axioms seem to be overly strong. Additional considerations when working with data from each cell are discussed below:

I. Data in this cell is the “best case” scenario. The scale generated via the Rasch model should have interval properties and all of the characteristics that make the Rasch model so attractive (e.g., specific objectivity). However, I expect that data will fall into this cell (at least in initial attempts to construct a measurement instrument) quite rarely.

II. With data from this cell, interval scales are possible but the most basic Rasch model (assuming that is what was used to scale the data) does not seem to fit. The state data would be an example. One has several options to choose from if data is in this cell. For example, more complex Rasch models may be used to try and improve fit, although one must be mindful of the fact that a more complex model may require the introduction of additional measurement axioms if interval scales are to be guaranteed. The most straightforward approaches here would
be to either rewrite or remove the items causing problems. One possible modeling approach for handling guessing (suggested in B. D. Wright, 1997) might also be worth considering. If guessing is the issue, a common lower bound for the item-response functions could be imposed. Since the item-response functions still can not cross in such a case (a possibility if the lower bounds differ from item to item), no additional axiomatic machinery is required. A more challenging technique that would allow for differential discriminations would be to attempt to use the model specified in Humphry (2011). One could also choose to abandon the parametric framework altogether and work with something like the doubly monotonic model of Mokken (1971). This could be a prudent decision for many applications since such nonparametric models tend to be more flexible than their parametric counterparts, but it is not likely to be taken up for the purposes of large-scale educational assessment due to their need for test equating, item banking, etc. Finally, one might consider other scaling methods (e.g., Kruskal, 1965) outside the traditional IRT framework. Whatever the decision, the key point is that if data is in this cell, then the work of creating an interval scale is not done.

III. The Duval, NC data from the Lexile assessment discussed in the previous chapter was an empirical example of data in this cell. Such data will require more than sophisticated modeling to produce an interval scale. If interval scales truly are needed, then I think further modifications of the measurement instrument itself must be considered. The results from the checks may suggest that certain characteristics of the items are at fault, in which case the measurement instrument can be refined. However, there are techniques available for such ordinal data (e.g., Johnson & Albert, 1999; Agresti, 2010) and these techniques may be adequate for many applications. If revising the measurement instrument seems overly costly, then perhaps such techniques are reasonable substitutes for those that require interval scales.

IV. Data that fits in this cell seems consistent with the Rasch model but not the axioms of ACM. This should not occur since the axioms of conjoint measurement should hold whenever the Rasch model fits. One way that data could be classified as belonging in this cell is if the data
are found to not be consistent with the axioms but, due to overly lax usage of fit statistics, the data is judged as fitting the Rasch model. To emphasize that data should not be placed in this cell, it has been struck out in Table 6.1.

6.1 Limitations

From a practical point of view, the most serious limitation of this work is that it only deals with dichotomously scored items and the Rasch model. One would naturally prefer to be able to easily accommodate scales that were based on polytomous items or scales constructed with more complex item response models for dichotomous data (such as the 3pl, Birnbaum, 1968) since these are frequently used in practice. However, incorporating either of these presents serious dilemmas. Let me first discuss attempting to incorporate more complex models for dichotomous responses. Polynomial conjoint measurement (Krantz et al., 1971) could be used as the basis for an argument that a scale created via the 2pl is interval. Kyngdon (2011) has demonstrated the plausibility of such an approach when discriminations are allowed to differ and Humphry (2011) is an alternative modeling approach that could be used to parametrize discrimination. But neither of these approaches helps with guessing. Since the vast majority of dichotomously scored data come from multiple choice items, this is a problem. One potential approach would be to simply fix a lower asymptote for all items (equal to the multiplicative inverse of the number of distractors). At least on tests where all of the items have the same number of choices, this would ensure that no item characteristic curves crossed due to guessing.\footnote{This suggestion is embedded in a series of “dos” and “don’t’s” for measurement in B. D. Wright (1997). The points are highly relevant for most educational testing scenarios. The mere fact that there was a need to make them in journal for practitioners, since most of his rules are constantly being broken, is perhaps a statement on the weak understanding held by most psychometrics professionals about what it means to measure.} However, such an approach may not be tenable. Guessing is perhaps a more complex phenomena that depends on attributes of the individual as well as the item. If it must be explicitly modeled, then arguments even more complex than those used in polynomial conjoint measurement are going to be needed to justify interval interpretations of such data.
Test designers include constructed response items since such items are believed to tap a broader range of cognitive processes than when multiple choice items alone are used. Scores to such items are frequently coded as polytomous rather than dichotomous since there may be a range of possible answers between which one can make multiple qualitative distinctions. For example, if an item asked the respondent to list the axioms of additive conjoint measurement, then a response which noted only single cancellation and double cancellation would deserve some credit but would be clearly be inferior to an answer that correctly named all four axioms. The rating scale model (Andrich, 1978) was the first complete development of a polytomous Rasch model (Rasch, 1960 and Andersen, 1977 had both begun work along these lines) but the partial credit model (G. Masters, 1982) is more common. If responses to item \( i \) have been coded in 0, 1, ..., \( m_i \), then this model posits that

\[
P(X_{ij} = x > 0) = \frac{\exp \sum_{k=1}^{x}(\beta_n - \tau_{ki})}{1 + \sum_{x=1}^{m_i} \exp \sum_{k=1}^{x}(\beta_n - \tau_{ki})}
\]

\[
P(X_{ij} = 0) = \frac{1}{1 + \sum_{x=1}^{m_i} \exp \sum_{k=1}^{x}(\beta_n - \tau_{ki})}.
\]

I include these equations to make the following point. Kyngdon (2011) argues that additive conjoint measurement for \( n \) components (Krantz et al., 1971) applies to the many-facet Rasch model (Linacre, 1989) since the quantities in the model follow the additive composition rule. The formulation of the partial credit model given here does not follow such a rule. Hence the natural analogy, additive conjoint measurement for \( n \) components, doesn’t apply and it is unclear what conjoint measurement framework could be used to argue that scales derived from polytomously coded responses are interval (and even more difficult still to argue that scales made up of both polytomous and dichotomous items are interval).

\[\text{3} \text{ Outside of educational assessment, Likert (1932) items that ask a respondent to gauge their attitude towards a question on a scale from strongly disagree to strongly agree are a common example of this type of item.}\]

\[\text{4} \text{ Note that in the denominator of each equation we would have sums of exponentials. These sums would not follow the additive composition rule.}\]
6.2 Future Research

There are two areas of further research that could be readily explored given the work done in this dissertation. This dissertation focused on inducing item-side distortions, but the same could be done with person-side distortions. Kyngdon (2008b) foresaw this possibility and even devised a basic set of research questions that one would want to answer. In particular, he wonders whether individuals whose responses support all axioms of conjoint measurement could be distinguished relative to those individuals whose axioms only support independence. He asks “would any of the person fit indices investigated by Karabatsos (2003) indicate individual violation or support of the conjoint measurement axioms?” (2008b, p. 103). I think it would be natural to examine the performance of the checks in such a context alongside person fit indices. It would also be interesting to know how much removing deviant respondents increased the overall quality of the data, whether judged by the checks or Rasch fit statistics.

My work on item-side distortions is not exhaustive. Depending on the nature of the test, distortions related to local dependence or multidimensionality may be of interest. In the Lexile data for example, dependence could be a culprit if cloze items are grouped into reading passages. Understanding whether dependence can be detected by the checks and, if so, whether it can be distinguished from alternative possible distortions, would be an interesting line of research. One could also compare the performance of the checks to the performance of item fit statistics specifically designed to detect dependence (e.g., van den Wollenberg, 1982; Wilson, 1988).

6.3 Implications for Practice

Considering the systematic nature of the scale distortions uncovered by the difference matrices, I think that greater attention should be paid to scale properties and the ramifications that misusing a scale might have, especially considering the increased reliance on educational assessments for the evaluation of not only students, but also teachers and schools. Techniques such as Betebenner (2009), that attempt to derive information even in the absence of an interval scale, are
one way of avoiding the dilemma. It would also make sense to probe further into how scale distortions may afflict specific test-based accountability designs that depend upon differences between scale values. These are large questions, but of potentially great importance.

Outside of the large-scale assessment context, the checks might be a useful tool in smaller applications where the Rasch model will be used for scaling. The checks can give a solid first indication of whether the Rasch model is appropriate. If so, all the better. If not, my hope is that practitioners will be inclined to think deeply about whether their questions require interval scales. If they do, then applying the Rasch model and mechanically producing estimates would be “business as usual”, but the underlying scale may be distorted. The first challenge in getting practitioners to use the methodology discussed in this dissertation is a practical one. I am currently in the process of creating a package for \texttt{R} (\textcite{R}) that will implement the methodology. This package, \texttt{ConjointChecks}, will take raw dichotomously scored data, produce the conjoint matrix where rows are grouped by sum scores, and then check a user-specified number of 3-matrices. This package will also offer the user the option of running the checks in parallel, an important feature since access to several computational nodes can reduce the computational burden substantially. Hopefully this package will make implementation of the checks by the user quite simple and lead to their use in diverse settings.

\section*{6.4 Closing Remarks}

Looking forward, it seems difficult to imagine that data from educational assessments won’t play an increasingly larger role in our educational accountability system. As such, it is natural to place the scales created from educational assessments under greater scrutiny. These scales will likely determine, in part, the fates of students, teachers, and schools. I want to close this discussion with a final example of the importance behind asking questions about scale properties.

The National Assessment of Educational Progress (NAEP) is frequently hailed as “the nation’s report card” and is the most prominent indicator of the quality of our nation’s primary and secondary schools. The Trend report (\textcite{NAEP}) documents
the long-term trends in the achievements of students at ages 9, 13, and 17. Consider the figures related to Reading from this report (the key numbers from this document—Figure A, page 3—are in Table 6.2). The fact that the difference between 1971 and 2008 scores is much larger for the age 9 students than the difference for age 17 students is frequently interpreted (Yglesias, November 1, 2011; Goldstein, November 1, 2011) as meaning that secondary schools have improved their reading instruction over the last 30 years less than elementary schools. This all depends, of course, on the quality of the scale. Given the difference in magnitudes, it is probably true that age 9 students have increased from 1971 to 2008 more than age 17 students. However, it would be a much bigger leap to assume that the growth over that period for age 9 students is a little over twice \((12/5)\) that of age 13 students. Such an inference requires an interval scale.

Table 6.2: NAEP Long-term trend results

<table>
<thead>
<tr>
<th>Year</th>
<th>Age 9</th>
<th>Age 13</th>
<th>Age 17</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>220</td>
<td>260</td>
<td>286</td>
</tr>
<tr>
<td>1971</td>
<td>208</td>
<td>255</td>
<td>285</td>
</tr>
<tr>
<td>Difference</td>
<td>12</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

The report goes on to describe the change in mean scores when disaggregated by race over the 1971 to 2008 time period. Since 1971, age 9 white students have increased 14 points. Black and Hispanic students have increased 34 and 25 points over the same time period. What are we to make of this? Interpreting these differences without reference to their starting point fundamentally

---

5 One potential scale-specific reason could be that the NAEP scale suffers from a so-called “ceiling effect.” In that event, it might not be that age 9 students have increased more.

6 A related point is made by Hoover (1984) who compares the growth from different points of the distribution of scores from two different scales. He notes that with one supposedly interval scale, “it is not possible to obtain the 1981 scale by a linear transformation of the 1973 scale. It follows that either the 1973 scale or the 1981 scale, or both, are not equal interval” (p. 12, emphasis in original). Due to the difficulty of creating interval scales, he suggests we use grade equivalent units instead.
requires an interval scale. While it would seem like the minority students have grown more than White students over that time period, any kind of quantification of this growth requires that the NAEP scale scores be on an interval scale. Designing a measurement instrument that has an interval scale is a non-trivial challenge and verifying that the scale is indeed interval should be a fundamental part of that design process.
References


Ferguson, A., Myers, C. S., R. J., R. J. B., Banister, H., Bartlett, F. C., Brown, W., et al. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. Advancement of Science, 1, 331–349.


Appendix A

Methodology Revisions

As I describe below, the initial revision of the methodology from Karabatsos (2001) was looser than it could have been. In particular, the check of double cancellation only restricted cell values based on one of two possible versions of double cancellation that must be checked within a 3-matrix. In this Appendix, I present an updated version of the methodology and compare it’s performance to the original version.

A.1 Double Cancellation, again

Double cancellation is most concisely stated within a 3-matrix as follows. If $x_1, x_2, x_3 \in X$ and $y_1, y_2, y_3 \in Y$ then for any assignment of $\{i, j, k\}$ to $\{1, 2, 3\}$ for the elements of both $X$ and $Y$, it must be true that

$$f(x_i, y_j) \geq f(x_j, y_i) \land f(x_j, y_k) \geq f(x_k, y_j) \Rightarrow f(x_i, y_k) \geq f(x_k, y_i).$$

Michell (1988) discusses that only 6 of the 36 possible double cancellation statements are logically independent of single cancellation, a point I leveraged in Chapter 3 when discussing the methodology. Assuming that rows and columns have already been ordered, this set of 6 can be easily identified using simplified notation from the following 3-matrix.

$$
\begin{array}{ccc}
  a & b & c \\
  d & e & f \\
  g & h & i \\
\end{array}
\quad (A.1)
$$
The 6 statements of double cancellation have to do with the 3 inequalities:

A : \( d < b \)

B : \( h < f \)

C : \( g < c \)

or their inverses (e.g., \(!A: b < d\)—note that I do not write \( b \leq d \) since the stochastic nature of the checks ensures that two quantities won’t be equal).

Ignoring equality, all possible scenarios relating to these 3 inequalities can be easily enumerated. I do this in Table A.1.\(^1\) I use “DC1” in that table to reference the version of double cancellation that I have been focused on thus far. Note that row 2 illustrates a violation of double cancellation. The problem is that row 7 is a violation of the second version of double cancellation to which my methodology would be insensitive. (Note that my rows 2 and 7 correspond to cell pairs 3 and 25 from Table 1 in Michell, 1988.) I refer to this second version of double cancellation as “DC2”.

It’s worth noting that the methodology of Karabatsos (2001) is similarly insensitive to this version of double cancellation since the \( c \) cell is forced to be larger than the \( g \) cell in the way that the jumping distribution is formed. Moreover, my oversight of this fact stemmed from the section of Michell (1988) discussing different version of double cancellation within a single cluster (such as DC1 and DC2 above): the double cancellation “instances are not all logically independent of each other. They fall into six cluster, the six instances within each cluster involving the same three pairs of cells (see Fig. 2) and being logically equivalent” (p. 471). DC1 and DC2 are within a single cluster and “logically equivalent” implies, in my reading, that they share the same truth table. Analysis of Table A.1 shows this is not the case.

\(^1\) It is illustrative to note that with probabilities simulated from the Rasch model, scenarios 1, 3–6, and 8 all occur roughly 1/6 of the time. Note that 2 and 7 should not occur since they would violate double cancellation and, indeed, they don’t.
Table A.1: Truth tables for double cancellation statements

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>A&amp;B</th>
<th>(A&amp;B) → C</th>
<th>!A&amp;!B</th>
<th>(!A&amp;!B) → !C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>4</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>7</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>
A.2 Updated Methodology

The changes in the methodology are relatively straightforward. For any of the non-corner cells, nothing in the updated approach is different. For the cells in the corner, new restrictions are imposed. Consider first the upper-right cell. There are 3 possible restrictions depending on the current values in the chains. These restrictions are still imposed stochastically based on the current values in the chain. Let \( \hat{\theta}_{(i,j)} \) be the current values in the chain. Let A represent the truth value of \( \hat{\theta}_{(2,1)} < \hat{\theta}_{(1,2)} \) and B the truth value of \( \hat{\theta}_{(3,2)} < \hat{\theta}_{(2,3)} \). If A and B do not have the same truth value, then no version of double cancellation needs to hold within this cell and the restrictions on \( \theta \) are

\[
\downarrow \theta \\
\uparrow \\
\downarrow \\
\uparrow
\]

If A and B are both true, then \( \theta \) needs to be greater than the value in the lower-left cell. The restrictions are then

\[
\downarrow \theta \\
\uparrow \\
\downarrow \\
\uparrow
\]

These two sets of restrictions were encoded in the first version of the checks. In the new version, if A and B are both false, then \( \theta \) needs to be less than the value in the lower-left cell. This restriction is

\[
\downarrow \theta \\
\uparrow \\
\uparrow \\
\uparrow
\]

A similar set of restrictions is placed on the lower-left hand cell.

It’s important to note that this updated methodology is more restrictive than the original methodology, although by how much is not yet known. What is clear is that data that seemed inconsistent with the axioms based on the checks (such as the data from the Lexile scale) should still be judged to be inconsistent with the newer version of the checks.
A.2.1 Technical Specification

The difference in the technical specification is only in terms of how \( l_3 \) and \( r_3 \) are defined. These two quantities are initialized as 0 and 1 respectively. Then, if \( i = 1 \) and \( j = 3 \) there are two possibilities:

- If the inequalities A and B from the above are both true, \( l_3 \) is set to \( \hat{\theta}_{(3,1)} \).
- If the inequalities are both false, then \( r_3 \) is set to \( \hat{\theta}_{(3,1)} \).

Note that in both cases the value that goes into \( l_3 \) or \( r_3 \) is the same, the question is whether it goes into the set of left- or right-hand limits. Similarly, if \( i = 3 \) and \( j = 1 \), then

- if A and B are both true \( r_3 \) is set to \( \hat{\theta}_{(1,3)} \) and
- if A and B are both false \( l_3 \) is set to \( \hat{\theta}_{(1,3)} \).

The second bullet in each of the two above lists is the restriction added in the updated version of the methodology.

A.3 Results

I compare the methodologies first via a brief example and then with simulated Rasch data.

A.3.1 A comparison of old and new methodologies

As a first comparison, posteriors were computed using the original and updated methodology for this 3-matrix:

\[
\begin{array}{ccc}
0.214 & 0.398 & 0.357 \\
0.449 & 0.463 & 0.456 \\
0.492 & 0.985 & 0.923 \\
\end{array}
\]

Note that the previously unused version of double cancellation is likely to play a role here since \( 0.398 < 0.449 \) and \( 0.456 < 0.985 \). In the updated version of the methodology, the great majority
of the checks involving the corner cells to involve additional restrictions (through either $l_3$ or $r_3$). In contrast, these restrictions are never used in the original methodology.

The additional stringency of the updated version of the methodology is almost undetectable, however. Figure A.1 compares the posteriors generated under each approach. The key point is that the 95% credible regions associated with the blue dots (based on the updated methodology) for the corners (near the dotted vertical lines) are tighter, but only very slightly so. This indicates that the methodology is unlikely to lead to qualitatively different results than the original version of the methodology in most cases.

**A.3.2 Simulation results**

To further demonstrate that the updated version of the checks perform similarly to the original version, I have recreated simple versions of the power and size analyses from Chapter 4. To do this, I again simulated data for 45 items (with difficulties evenly spaced between -1.5 and 1.5) and 15,000 respondents. In the purely Rasch version, 2% of the checks detected violations although this dropped to 1% when the violations are weighted by the number of respondents per sum score. These are almost identical results to those from Table 4.1 in Chapter 4.

In the next step, I modified the above simulation only slightly by first introducing a single item to have a non-zero guessing parameter. I then modified a single item to have a non-unity discrimination parameter (in this second stage, the non-zero guessing parameter from the previous step has been set back to zero). Figure A.2 demonstrates that deviant items are still easy to identify visually. Each item has 13% of its checks detecting violations. In each case, the item with the next most violations has fewer than 5% of its checks detecting violations. Based on this evidence, the checks can still detect deviant items.
A.4 Why the updated methodology doesn't affect the key results from the dissertation

In the above section, I demonstrated very comparable results between the two methods in a variety of situations. I think there are additional reasons above and beyond what I showed there to trust the results that utilize the checks in the body of this dissertation. First, my approach was entirely normative. If I had been, for example, relying on a specific percentage of violations then this would be a problem since the updated methodology may change the number of violations detected. Conclusions were based on comparing the percentage of detected violations via the original methodology for a given data set to those detected in a known, purely Rasch data set. The second main point would be that since the main conclusions of the empirical section are that the empirical data generated violations at a higher rate than the simulated Rasch data, more stringent tests would only make this worse.
Figure A.1: Comparison of 95% credible region for posterior from updated and original methodology. Colored circles represent mean of posterior under either method. Larger circles represent observed proportions of correct responses for the 3-matrix (note the violation for cell (3,2)). Dotted vertical lines indicate “corner” cells.
Figure A.2: Demonstration of ability of updated methodology to detect items that deviate from Rasch model via either a non-unity discrimination or non-zero guessing.

Increasing Ability

Item in red has 0.3 Guessing Parameter

Increasing Ability

Item in red has 1.5 Discrimination Parameter