NEW TOOLS FOR INTERVENTION FIDELITY ASSESSMENT:

AN EMPIRICAL COMPARISON OF EXPLANATORY MULTIDIMENSIONAL IRT AND CTT APPROACHES

By

Michael Cader Nelson

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Psychology

May, 2013

Nashville, Tennessee

Approved:

David S. Cordray, Ph.D.

Sun-Joo Cho, Ph.D.

Andrew Tomarken, Ph.D.

Bethany Rittle-Johnson, Ph.D.

To Marcie, who solves the unsolvable problems.

ACKNOWLEDGEMENTS

This work is only possible through the financial support of the Vanderbilt University Graduate School, as well as a pre-doctoral fellowship in the Experimental Education Research Training (ExpERT) program from the Institute of Education Sciences (#R305B04110).

This paper reflects all that my advisor, Dr. David Cordray, has taught me, both by instruction and by example. Thank you for pushing me beyond numbers toward ideas, beyond answers toward questions, and beyond repeating conventions to challenging them, particularly my own. I am also thankful to my collaborators and colleagues whom I learned beside and from, including Dr. Chris Hulleman, Evan Sommer, Dr. Catherine Darrow, and Dr. Amy Cassata.

One measure of an excellent teacher is that you continue learning from him or her even years after instruction, and every member of my committee meets this criterion: I continue to reference old notes and texts, am reminded of advice and admonitions, and in some cases learn indirectly through your advisees. I particularly want to acknowledge the patience and generosity of Dr. Sun-Joo Cho, attributes which have been equaled only by her knowledge and ingenuity.

I also want to thank my friends, from the very old ones who long endured my peculiar fascination with puzzles, language, the mind, and the vigorous debates that emerge at their intersection; to my newer friends who stoke these passions daily. Cori and Michele, you especially have supported me in my victories and defeats, and allowed me to share in your own.

Finally, Marcie, you know you have my unending and immeasurable love and gratitude. Half of the credit for all my work goes to you, and more than half in states without community property laws.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

NEW TOOLS FOR INTERVENTION FIDELITY ASSESSMENT:

AN EMPIRICAL COMPARISON OF EXPLANATORY MULTIDIMENSIONAL IRT AND CTT APPROACHES

Intervention fidelity (Nelson, Cordray, Hulleman, Darrow, & Sommer, in press) is the extent to which an intervention has been implemented as planned in the treatment group, and differentiated from the control group, in the context of a randomized controlled experiment (RCT). Education researchers are seeking more and better tools for measuring intervention fidelity, but approaches have varied widely among researchers, and there are few direct comparisons of different analytical methods.

IRT approaches may be especially capable of overcoming difficulties associated with analyzing intervention fidelity data, including skewed distributions, multidimensionality, and poorly-defined constructs. A recent development in explanatory multidimensional IRT (MIRT) is a model that detects group differences and individual differences simultaneously for multidimensional tests (Cho, Athay, & Preacher, in press). Model results then can be compared directly with factorial analysis of variance (ANOVA, Kirk, 1968) results.

The primary goal of this study is to demonstrate parallel analyses of empirical intervention fidelity data with both the traditional ANOVA using total scores and this particular MIRT model. The comparison shows the unique strengths of explanatory MIRT for intervention fidelity analyses, allowing researchers to assess its benefits over classical test theory (CTT) approaches, as well as the feasibility of MIRT analysis for their data. Secondarily, the results of this study show that choice of analytical method for intervention fidelity analysis can lead to

somewhat different statistical conclusions. It is recommended that the sources of such deviations be investigated through simulation studies and other methods in the future.

Introduction

There is a growing recognition among researchers and evaluators that fully understanding the outcomes we measure and interpret also requires measuring and interpreting the extent to which the causes of outcomes have been implemented as intended. This is the most basic definition of fidelity of implementation, though many elaborations have been developed across investigators, studies, and disciplines (O'Donnell, 2008). Fidelity is commonly assessed in a number of fields (e.g., education, medicine, behavioral health, community programs) because interventions are less likely to provide the greatest benefits (or perhaps any benefit at all) if not fully implemented (Durlak & DuPre, 2008; Dusenbury, Brannigan, Falco, & Hansen, 2003; Harachi, Abbott, Catalano, Haggerty, & Fleming, 1998).

Fidelity is particularly important in the context of randomized controlled trials (RCTs) for testing the intervention as intended, and for obtaining the maximum contrast between treatment and control groups to measure treatment effects. Assessing fidelity helps researchers to make valid conclusions about the processes that cause experimental outcomes: it opens up the experimental "black box" to expose intervention mechanisms in practice, thereby supporting or rebutting theorized conceptions of the cause (Cordray & Pion, 2006; Harachi et al., 1999). Failing to assess fidelity can lead researchers to conclude inaccurately that the intervention as intended achieved or did not achieve the measured outcomes (Dusenbury et al., 2003; Harachi et al., 1999), so that an effective intervention might be discredited or an ineffective one supported.

The term "intervention fidelity" has been used (Nelson, Cordray, Hulleman, Darrow, & Sommer, in press) to describe a subtype of fidelity distinguished by being 1) assessed in the context of an RCT, 2) model-based, and 3) focused only on intervention core components and the extent to which these are present in the treatment condition and absent from the control condition. Focusing on intervention fidelity narrows the scope of what otherwise is a diverse set of theories and analytical tools (O'Donnell, 2008), while relating the (typically correlational) assessment of causes to Rubin's causal model for measuring effects (Holland, 1986). Intervention fidelity is model-based in that the intervention's intended processes are conceptualized initially as causal constructs (Chen, 1990; Knowlton & Phillips, 2009) and then operationalized in a logic model (Kellogg, 2004; Knowlton & Phillips, 2009). The model specifies treatment core components, how those components affect one another, how they differ from control conditions, and when to measure those differences.

Assessing Intervention Fidelity

Education researchers frequently approach methodologists with some version of the following question: "I know that fidelity is important and I want to measure it and link it to my outcomes, but how do I go about doing that?" Cordray & Pion (2006) answered this question in part by proposing a systematic procedure for assessing intervention fidelity, later elaborated by Nelson and colleagues (in press): 1) specify the intervention model, 2) identify indices, 3) determine reliability and validity, 4) combine indices as appropriate, and 5) analyze the difference in fidelity between conditions, linking to outcomes where possible. Yet steps two through five are undermined by the lack of proven tools and procedures for measuring and analyzing intervention fidelity in the literature (O'Donnell, 2008). As researchers seek best practices for the relatively new practice of rigorous fidelity assessment in RCT's, it is unclear

that traditional methods for analyzing outcomes have the same strengths and weaknesses for analyzing intervention fidelity.

In this context, item response theory (IRT) methods may be especially advantageous for assessing intervention fidelity. IRT has been touted as a providing a systematic means of developing instrumentation that is both reliable and valid, as well as analytical methods that are superior to those used in Classical Test Theory (CTT).

The Promise of Item Response Theory

In recent decades, IRT's advocates have championed it as establishing "the new rules of measurement" (Embretson, 1996), rendering CTT "obsolete" (Zickar & Broadfoot, 2009, p. 37). They urge the adoption of IRT methods in social science fields outside of ability testing, where it emerged and now dominates. Others, however, have objected to the predicted demise of CTT as "shortsighted" and "premature," an "urban legend" and a "myth" (though one with a "kernel of truth") (Zickar & Broadfoot, 2009, pp. 37-38). Methods from the two theories have been compared in numerous studies, simulated and empirical (Reckase, 1997), resulting in mixed conclusions: most simulations have demonstrated that CTT results can present problems that IRT resolves (c.f. Embretson, 1994, 1996), while other simulations and most empirical studies largely obtained equivalent results for the two methods (c.f. Fan, 1998; MacDonald & Paunonen, 2002; Osteen, 2010; Amin, 2011; Willse & Goodman, 2008; Ndalichako & Rogers, 1997).

IRT and CTT methods have not previously been compared for assessing intervention fidelity; in fact, a thorough search of the literature yielded only one study—a conference paper (Williams, Swanlund, Miller, Konstantopoulos, & van der Ploeg, 2012)—that employed IRT in its analysis of fidelity within an experiment. In fact, IRT is only beginning to be employed

beyond the context of ability testing (Reise & Henson, 2003) and rarely to randomized experiments in the social sciences (although, see: Jago, Baranowski, Baranowski, Thompson, Cullen, Watson, & Liu, 2006; Baranowski, Allen, Masse, & Wilson, 2006; Watson, Baranowski, & Thompson, 2006). IRT is not likely to be embraced fully in fields outside of ability testing until there is "research that demonstrates that the use of CTT methods can lead to incorrect substantive conclusions, whereas an IRT approach leads to more valid substantive findings" in those fields (Reise & Henson, 2003, p. 101).

In the following sections, I describe general model characteristics and assumptions for both CTT and IRT. Much more comprehensive descriptions of each approach can be found elsewhere (c.f. Lord & Novick, 1968, for CTT, and Embretson & Hershberger, 1999, for IRT). This general discussion is followed by more specific comparisons of the theories in the context of assessing intervention fidelity.

Classical Test Theory

In general, the CTT model conceptualizes observed test scores as a linear combination of respondent true score and error score, as shown in equation 1:

$$X_j = T_j + E_j \tag{1}$$

where X is the observed score, T is the true score, and E is the error score, with subscript j for person. The true score is not an inherent trait of the individual alone, but is dependent on the measure: a particular test should yield the same true score with varying error for a single

individual (assuming no ability change), but that individual may have different true scores for different tests of the same ability (Hambleton & Jones, 1993).

In the context of analysis of variance (ANOVA) models, CTT has three assumptions: 1) true scores and error scores are uncorrelated, 2) the error in the respondent population averages to zero, and 3) parallel forms of a test have uncorrelated errors. Parallel forms measure the same trait with different items, and yield the same true scores with equal measurement errors. Other CTT approaches are derived by revising or adding to these assumptions.

Test reliability in CTT is defined as the proportion of observed score variance that is attributable to true score variance, and is operationalized and measured in several ways, e.g., Cronbach's alpha (1951) measures internal consistency. The standard error of measurement (SEM) is the expected value of the error and can be used to construct confidence intervals.

The chief advantages of CTT are its weak assumptions and its familiarity. Criticisms include that scores are usually test-based and not item-based, scale characteristics are sample dependent and scores are scale-dependent (i.e., circular dependence), and that it reduces interval-level variables to ordinal-level measurements (Embretson & Hershberger, 1999). Interval scaling is violated because individual scores are dependent on item properties (e.g., different items have different levels of precision), impacting the overall score distribution and parameter estimates (Embretson, 1996). Technically, it is possible for CTT to achieve interval scores under narrow assumptions (Embretson & Reise, 2000): that the scale is measuring interval-level *true scores* that are normally distributed *in the population*[1] (Jones, 1971). If these assumptions are met, an

[1] As opposed to *observed scores* being normally distributed *in the sample*, where the central limit theorem can achieve statistical normality. Jones (1971) gives the hypothetical example of intelligence being conceived of as resulting from such a large number of independent influences that that the central limit theorem would apply to the parameter itself.

interval scale results from the linear relationships among item scores relative to the trait (Embretson & Reise, 2000) and between distributions of test scores when comparing groups (Yen, 1986). Even if it were possible to verify the attributes of true scores independent of error, measurement would only be guaranteed for samples from the test's norming population.

Some might suggest that factor analysis (FA) and its derivative, structural equation modeling (SEM), should be considered CTT alternatives to IRT because they describe latent constructs (Willse & Goodman, 2008) and are more familiar and accessible for certain applications (Zickar & Broadfoot, 2009). Indeed, both unidimensional and multidimensional IRT can be derived from a subset of FA models (Kamata & Bauer, 2008; Reckase, 1997), and SEM can be used to compare latent scores across groups and can provide a standardized effect size (Willse & Goodman, 2008). For example, Willse and Goodman's (2008) simulation study compared SEM effect sizes estimated in the multiple indicator multiple cause (MIMIC; Zellner, 1970) framework, raw score standardized group differences with Cohen's *d*, and IRT "effect sizes" calculated as the difference between group trait level estimates divided by their pooled standard deviation. Findings showed little difference among the results of the three methods for estimating effects when applied to the same simulated data.

Such comparisons, however, can be misleading: FA and SEM methods can be used for either CTT or IRT analyses of data, depending on the measurement model. Categorical item factor analysis is mathematically equivalent to IRT (Kamata & Bauer, 2008), possibly explaining why some comparisons of the two methods (Amin, 2011; Osteen, 2010) produced similar results between models. Likewise, SEM can be equivalent to IRT if the manifest variables in the measurement model are categorical item responses rather than CTT test scores. In such instances, it is not meaningful to claim that FA or SEM is as good as or preferable to IRT

methods precisely because they effectively *are* IRT methods. In other instances, when CTT test scores (not item-level responses) are analyzed using FA or SEM, the latent trait level estimates continue to be vulnerable to violations of CTT assumptions. Also, like IRT, the quality of SEM results and interpretations depend on correctly specifying the model, and may also be susceptible to violations of multivariate normality depending on the estimation method used (Boomsma & Hoogland, 2001).

Item Response Theory

In contrast to CTT, IRT models aim to estimate theta (θ), a latent individual ability or trait that is independent of the scale by which it is measured. Instead, both trait level and item difficulty (β) are estimated on the same scale, such that any person with a particular trait level theta will have 50% chance of correctly answering any relevant, unidimensional item of equal difficulty (β). If item properties influence scores above and beyond theta, this would violate the IRT assumption of unidimensionality.[2]

The IRT one parameter logistic (1PL) or Rasch model can be represented as equation 2:

$$\text{logit}\left[P\left(y_{ji} = 1|\theta_j\right)\right] = \left(\theta_j - \beta_i\right) \quad (2)$$

where y_{ji} is the response, is θ_j is the latent person parameter and β_i is an item location parameter (difficulty). Different IRT models can be formed by adding parameters (item discrimination for 2

[2] This is absolutely true of theta but not necessarily for the estimate of theta, because many tests will have some level of multidimensionality (Zickar & Broadfoot, 2009). Small amounts of multidimensionality do not interfere with parameter estimation (Reckase, 1979). Influential item characteristics can be detected through several procedures (see differential item functioning [DIF] detection, below) and can be accounted for in the analysis.

parameter logistic, discrimination and guessing for 3parameter logistic), allowing for different kinds of responses (binary or polytomous), accounting for multidimensionality, and linking item responses to person estimates using a particular functional form (e.g., cumulative logistic distribution). All IRT models share the same strong assumptions that the latent trait is unidimensional and that items are uncorrelated except through the latent person parameter (i.e., local independence).

Reliability in IRT is related to information, such that items and tests have the most precision and provide the most information where difficulties match ability levels in Rasch family models. Test and item information curves allow one to identify for whom a test will produce the most precise and informative measurements. The model fit to the data, or separation reliability, is defined as the proportion of variance in responses explained. In addition to their information, items also can be described in terms of the item characteristic curve (ICC), an ogive for which the slope, position relative to the ability scale, and probability of a correct response are determined respectively by item discrimination, difficulty, and guessing parameters (Embretson & Hershberger, 1999).

The main advantages of IRT measures are that they are population invariant, scale invariant and interval level. This is possible because the latent person parameter is estimated from relevant item properties (e.g., difficulty), placing both trait and difficulty levels on the same interval scale (Embretson, 1996).[3] In the Rasch model, for example, responses are a simple linear combination of item difficulty level and person trait level (all in log odds units; Embretson & Reise, 2000). Misconceptions about IRT score characteristics may be responsible for some of the

[3] More technically, it can be shown (Fischer & Molenaar, 1995) that this follows from the principles of local stochastic independence, the sufficiency of raw scores for characterizing the latent person parameter, and the definition of the relationship between traits and items by a continuous, increasing, strictly monotonic function $f(\theta - \beta)$, where β is a real number.

mixed empirical results in earlier comparisons with CTT: IRT ability estimates frequently are compared with CTT scores through correlation (Fan, 1998; MacDonald & Paunonen, 2002; Ndalichako & Rogers, 1997), which identifies ordinal but not interval similarity. Other studies have compared IRT and CTT parameter estimates or effect sizes without first establishing true comparability (Fan, 1998; Ndalichako & Rogers, 1997; Willse & Goodman, 2008).

The main disadvantage of IRT generally is that its assumptions, while few, can be difficult to meet for some models and with types of some data. Worse, IRT methods are not robust to violations of these assumptions. In many cases, IRT analyses also require software that is not widely available and that requires special knowledge to use and to interpret results.

Can IRT provide superior methods for intervene fidelity assessment?

Though previous studies have compared IRT and CTT methods for assessing ability levels and experimental outcomes, the two approaches have yet to be evaluated specifically for fidelity assessment, which has different characteristics and raises different concerns. Described below are four ways IRT might better address fidelity-specific concerns.[4]

1. Latent constructs versus manifest variables. Some would argue that fidelity of implementation should be conceptualized as a purely manifest variable, merely the sum of the indicators measured, called an "emergent construct" (Reise and Henson, 2003). However, in the context of intervention fidelity, the core components of the intervention are specified in terms of latent constructs that ultimately cause the treatment effects. The extent to which a person's behaviors manifest the intended intervention constructs, then, can be estimated as the latent

[4] This discussion assumes that the fidelity indices were not developed using IRT methods, which involve another set of advantages and disadvantages.

person parameter, the latent person trait that probabilistically causes item responses (Willse & Goodman, 2008; Reise & Henson, 2003).

CTT raw score approaches use observed scores to estimate an individual's true score.[5] Items on a fidelity scale tend to be common educational activities that the intervention design has linked in a particular way to promote learning, but items must be kept generalized when the scale has been designed to measure fidelity across conditions (Nelson et al., in press). CTT raw scores, though typically highly correlated with IRT latent scores (MacDonald & Paunonen, 2002), may be unable to discern which items are only superficially similar to the underlying construct or add little additional information about it. As a result, fidelity levels may be overestimated or underestimated by CTT scoring. From the IRT perspective, responses to an irrelevant item cannot be predicted from the overall estimate of the latent person parameter, which would be indicated by item fit statistics. A set of such items, once identified, can be excluded or accounted for through multidimensional IRT models. Likewise, IRT methods can identify items that may measure the construct of interest but add little additional information for scoring.

Furthermore, the constructs of intervention processes often are more ambiguous than those underlying the main intervention outcomes. The intervention may have been designed without ever explicitly articulating its components except on a procedural level, the underlying constructs remaining implicit. In fact, it is not uncommon for the designer, implementer, and evaluator of the same intervention each to have a different understanding of *why* the intervention would work, much less how to measure it, and never even realize it (Chen, 2005). As a result, measures may include items that are less related to the target construct that will be weighted

[5] The latent person parameter is actually a different variable than either the observed score, which is how the person responded to the test, or true score, which is how the person would have responded to the test without measurement error (Hambleton & Jones, 1993).

equally by CTT analyses without theoretical justification or empirical evidence (unless additional procedures like factor analysis are employed). Descriptive IRT methods attempt to match item difficulties with a priori construct levels, and evidence for or against the theoretical framework is obtained by analyzing the extent to which actual responses match expectations. Explanatory IRT methods, which seek to account for variance in the data even without clear construct mapping, automatically evaluates appropriate item contributions to overall scores by estimating how much information each item provides about the common latent construct (Fraley, Waller, & Brennan, 2000).

2. Multidimensionality. Of course, researchers employing IRT methods will fail to correctly estimate the latent person parameter unless it meets the assumption of unidimensionality. This is likely to be an issue in assessing fidelity: complex interventions involve implementing a chain of constructs, with individual constructs often conceptualized as having multiple subconstructs. It may be useful in such cases to measure basic components separately before creating a composite. As such, the intervention scales may be multidimensional, or the items used to assess them may be sensitive to other, unrelated constructs. Using IRT methods to analyze these scales will lead to model misfit and biased scores, and because IRT is item-based, multidimensionality is an issue for each item. Some even hypothesize that interventions could cause multidimensionality between experimental groups, although there is little empirical evidence of this (Baranowski et al., 2006).

Fortunately, there are numerous IRT methods that can account for multidimensionality. There are several statistical methods for identifying items sensitive to nuisance dimensions and biased scores for one subgroup of respondents, called differential item functioning (DIF) (Millsap & Everson, 1993). Once detected, these items can be revised or replaced (during

instrument development) or omitted entirely if doing so does not undermine scale reliability and validity. When the construct of interest is theorized to be multidimensional, data can be fitted using multidimensional item response theory (MIRT) models, preserving the assumption of unidimensionality and even improving precision and reliability for shorter tests (Wang, Chen, & Cheng, 2004). Comparing unidimensional and MIRT model fit assesses how well the data conform to the theorized structure (Rijman, 2010; Yao, in press).

The requirement of assessing dimensionality actually may be an advantage for IRT, given that multidimensionality often is unaccounted for in CTT analyses. Simply combining raw data across dimensions implicitly weights the dimensions (Ackerman, 1992) and confounds score interpretations (Zickar & Broadfoot, 2009). Applying SEM and FA to total scores are options for assessing dimensionality of complex constructs measured with multiple scales, and the software for these are more widely available and taught than IRT software. Though useful, these methods do not assess item difficulty and error (which would by definition be within the IRT framework) and must also meet the assumptions of CTT (Reckase, 1997; Zickar & Broadfoot, 2009).

3. Other test and item characteristics. Even when measured with precision, the distribution of fidelity levels among respondents presents a challenge: in the ideal case of perfect fidelity in the treatment group and total differentiation from the control group, there would be no within-group variation and so no distributions to test. Even in the more realistic context of very high treatment fidelity and very high contrast between groups, one would expect the two groups' distributions to be highly skewed in opposite directions. Non-normal distributions violate a basic CTT assumption, and performing transformations or Winsorizing outliers may obscure the defining attributes of fidelity one is seeking to detect. Though the ideal case is unrealistic, many studies achieve near-perfect fidelity because, unlike in other areas of assessment, researchers

may actively intervene during the data-collection process when the desired outcome is not being obtained. Even when fully-developed interventions are implemented in the field, they often incorporate fidelity drivers (Fixsen et al, 2005) like manualization, training, and ongoing coaching and feedback. Through such means, Peer Assisted Learning Strategies (PALS) has achieved fidelity levels around 90% across decades of studies (Fuchs, Fuchs, & Burnish, 2000).

Normality of distributions is not an assumption of IRT, and item characteristics are prevented from influencing characteristics of the distribution: items' individual properties (e.g., difficulty and precision) are incorporated in IRT models, allowing one to account for these properties when estimating the latent person parameter (Embretson, 1996). Also, while CTT scales may have different properties for different populations (e.g., ceiling and floor effects), the IRT property of item invariance allows one to assess any population using non-DIF items. Thus, one may use the same instrument to measure fidelity within the control group as well as within the treatment group that (with an effective intervention) diverges from it (Embretson, 1996; Hambleton & Jones, 1993; MacDonald & Paunonen, 2002). It is not even necessary to standardize differences in the same ability between groups, because their means are on the same logit scale. CTT scores for the same ability must be converted to a common scale for comparison, and the same person is expected to have different scores on different tests (Embretson, 1996).

4. Sample sizes. A final point can only be characterized as a weakness of IRT compared to CTT, although it can be minimized. IRT analyses typically require larger samples than CTT analyses because IRT models generally contain more parameters (for items and individual scores) and, in general, standard errors for parameter estimates increase as the number of respondents and items decrease (Zickar & Broadfoot, 2009). Unfortunately, even large-scale

education experiments can be small relative to the usual sample sizes for ability testing, which often are in the thousands. The sample is further reduced for intervention fidelity if it is only measured within the treatment group. Worse, because fidelity assessment is sometimes a secondary consideration and resources are scarce, the sample size may be reduced again if fidelity data are only collected from a subsample of the unit of analysis (teachers, students, etc).

The sample size requirements for precise IRT parameter estimation can be reduced under certain circumstances. It is true that recommended cutoffs for adequate IRT sample sizes have been as high as 500 (Embretson and Reise, 2000), estimating parameters with one-parameter models requires fewer responses (Zickar & Broadfoot, 2009), and Linacre (1994) suggested sample sizes of as few as 50 to 100 when applying Rasch models. When MIRT models are appropriate, its precision can be greater with smaller samples by virtue of using information both within and between the multiple constructs (Wu & Adams, 2006; Yao, in press).

The Current Study

The aim of this study is to compare and contrast particular IRT and CTT methods for the analysis of fidelity between subgroups of experimental participants. I will conduct separate analyses of empirical intervention fidelity data from a large-scale RCT with the two-way between-groups ANOVA (Kirk,1968; Maxwell & Delaney, 2004) CTT procedure and an IRT procedure employing a version of the general explanatory longitudinal item response model (Cho, Athay, & Preacher, in press). Each analysis will test whether teachers' implementation of differentiated instruction differs between treatment and control conditions or between grade levels, and if there is an interaction between experimental condition and grade level. It is not possible to know the "correct" parameter values for empirical data, but finding that the two

models yield results that lead to different statistical conclusions about fidelity would constitute evidence that the theoretical advantages between IRT and CTT may be of practical importance.

The research question for this study is: Will the IRT and CTT analyses of the same intervention fidelity data yield different statistical conclusions from one another about the population differences among experimental conditions, grades, or their interactions? If so, I will compare the two analyses to identify possible causes. Finally, I will discuss the implications for choice of analytic tools for fidelity assessment.

The methodology of this study represents an advantage over many of the previous attempts to compare IRT and CTT using empirical data. Many of the studies referenced above directly compared parameter estimates or effect sizes between methods, a procedure with arguable validity at best given the need to convert IRT parameter estimates to an ordinal scale for comparison with CTT parameter estimates. The present study instead compares parameters for separate groups within each method, then compares across methods on the results of their respective significance tests (reject or fail to reject the null hypothesis). While not conclusive as to which approach is superior, the findings of this study can serve as the foundation for building a body of evidence as to the usefulness of IRT methods for assessing intervention fidelity.

Methods

The MAP intervention

Intervention fidelity can only be studied empirically in the context of an RCT for a particular intervention. The data for the present analysis are from an efficacy study (Cordray,

Pion, Dawson, & Brandt, 2008) of the Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) reading intervention. In brief, MAP provides teachers with tools for differentiating reading instruction in their classrooms. MAP training and tools were designed to improve student reading by helping teachers to implement differentiated instruction (DI) more effectively, which has been shown to improve student outcomes in diverse classrooms (NWEA, 2003).

The MAP tools for teachers include specially-designed computer software that facilitates regular assessment of individual students' reading skills. Teachers who receive training and coaching in the MAP program are expected to use MAP software to help structure DI in their classrooms. A highly simplified model for the MAP intervention is depicted in Figure 1.

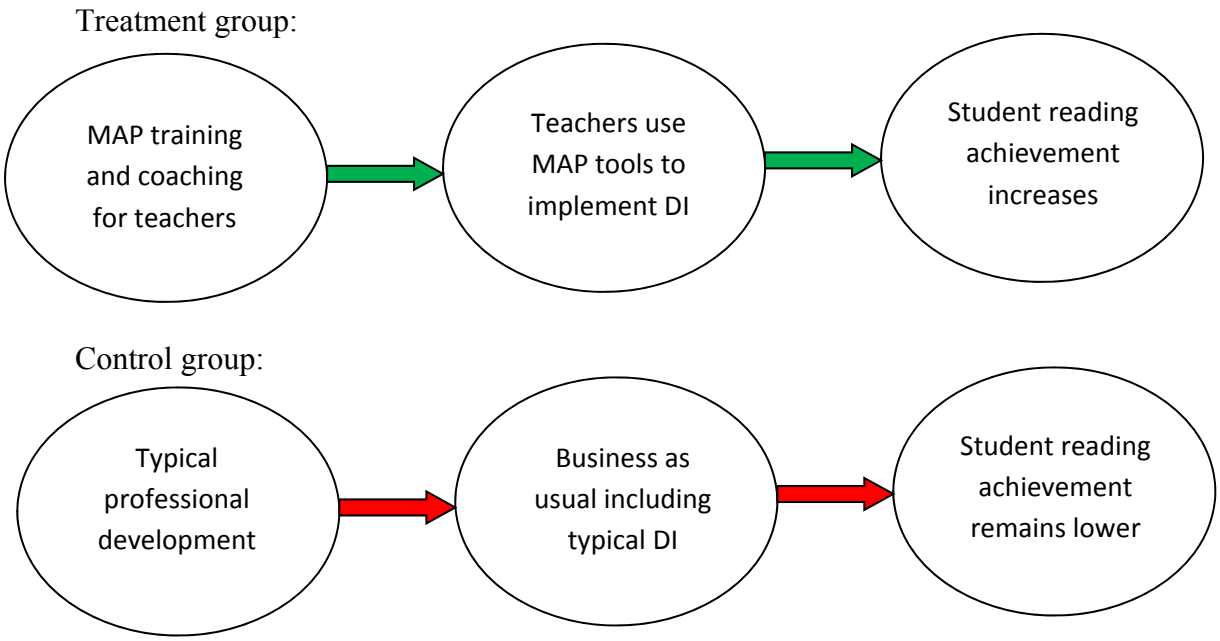


Figure 1. Simplified MAP intervention model (Cordray et al., 2010)

The second component in Figure 1 constitutes the unique MAP process of implementing DI. Yet this process is only a particular (though hopefully highly effective) operationalization of the general construct of DI.

Ideally, intervention fidelity is measured in each experimental condition and the difference between conditions (the first two components in Figure 1) is considered the cause of intervention effects (the last component in Figure 1) (Hulleman and Cordray, 2009; Nelson et al., in press). However, program-specific fidelity usually is assessed only within the treatment group because unique program elements are unlikely to be found in the control condition except in the most extreme cases of contamination. Instead, comparing intervention fidelity between conditions requires measuring intervention components at the construct level.

The DI construct

The key MAP intervention construct of DI is likely to occur in a business-as-usual contrast condition because DI has long been understood to be an effective tool and encouraged as a standard best practice (Tomlinson and McTighe, 2006). In fact, teachers receive both preservice and inservice training and support for DI in reading and other subjects. DI occurs whenever an educator uses benchmarking data of students' knowledge and ability levels to fit instruction to individual students' needs, and often involves ability grouping using student performance data (Tomlinson, 2001; Hall, 2002). As diversity increases in classrooms, teachers are encouraged to adopt DI practices as one strategy for addressing a variety of interests, learning styles and actual knowledge (Tomlinson and McTighe, 2006).

The ambiguity of the DI definition often leads to DI practices that are idiosyncratic to a particular classroom, depending upon several situational factors, including the availability of a

valid source of student data (Decker, 2003). In short, teachers seeking to implement DI should benefit from a model of effective practices and tools they can employ. The MAP intervention is an attempt to provide such a model. RCT researchers worked with the MAP developers to understand and describe DI as embodied by the MAP framework (Cordray et al., 2010).

The MAP DI construct is theorized to involve three interrelated subconstructs (Cordray et al., 2011): teachers must form instructional groups according to students' reading ability (instructional modality or IM), use instructional strategies appropriate for each group (diverse instructional strategies or DS), and provide content on topics appropriate for each group (instructional topics or IT). However, the DI construct is more than merely the sum of these manifest activities. For example, using diverse strategies like allowing students to work in different locations in the room or in other areas of the school (both items in the IS subscale) potentially could help facilitate DI, but they are often merely a function of logistics, and teachers may not take advantage of these situations to promote DI. When the three subconstructs are implemented for the purpose of individualizing and optimizing the educational experience, DI is the shared element among them. This relationship is portrayed in Figure 2.

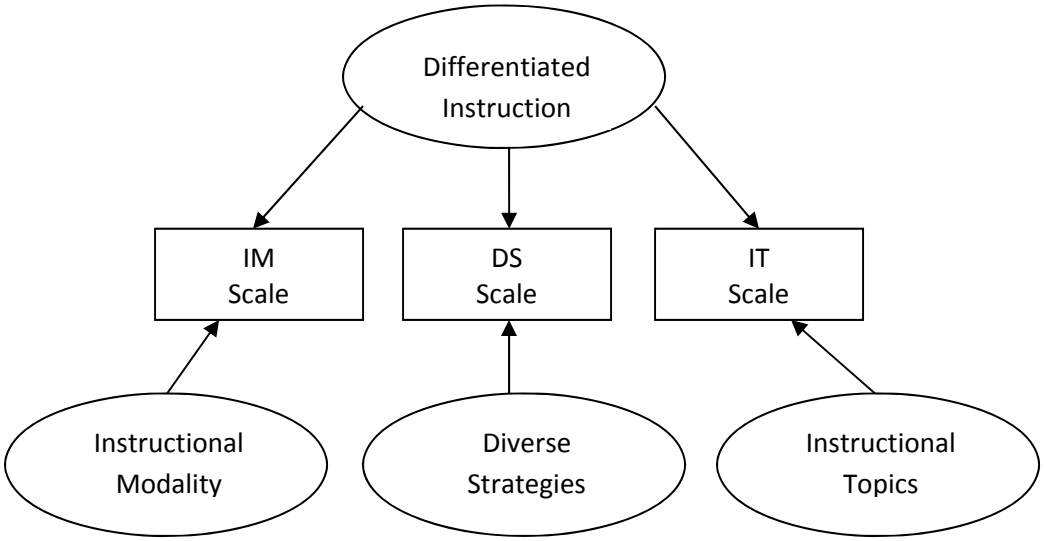


Figure 2. A representation of the DI construct as defined in the MAP study

The extent of DI implementation may be influenced by teacher characteristics. In addition to any effects of treatment, DI could vary between grade levels, perhaps due to differences in student needs, curriculum, or teacher beliefs.

Data Sources and Variables

The data for this analysis were collected during the first year of a two-year RCT, *The Impact of the Measures of Academic Progress (MAP) on Differentiated Instruction and Student Achievement* (Cordray, et al, 2008), funded by the Institute of Education Sciences (IES). To assess the effectiveness of the MAP program for elementary reading, 32 schools in five Illinois districts were recruited, and the 178 fourth and fifth-grade teachers were randomly assigned at the grade level within-school to the MAP treatment or to the business-as-usual control condition. The study was powered to detect an effect size of $g = 0.2$, but low or no effects were anticipated in the first year due to a late start.

Implementation data were collected in part using the year-end *Survey of Teachers on Instructional Practices* (Cordray, et al, 2010), which was developed specifically for this study using select items from the *Teacher Questionnaire* (Study of Instructional Improvement or SII, 2001) and Section III of the *Survey of Enacted Curriculum* (Blank, Porter, & Smithson, 2001). The relevant subscales (see Appendix A) were designed to assess general construct-level fidelity in each of the three DI domains, and surveys were administered to teachers in both conditions. 168 teachers responded (94%), and missing data within these responses were imputed using sequential regression multiple imputation (SRMI) (Raghunathan et al., 2001). Survey data were obtained and processed as approved by the Institutional Review Board of Vanderbilt University (Cordray et al., 2008), and de-identified prior to the current analysis.

Most items were designed for Likert-scale responses describing the frequency of activities in each of the three DI domains, but responses were dichotomized based on implementation thresholds developed by MAP designers and evaluators. Dichotomizing responses is consistent with the DI construct and the notion that it is either implemented as intended or not. The DS and IT subscales each had two sections with the same items but referencing students who were either high-level or low-level readers. Teachers' responses within these sections were scored by whether they were appropriate for students of the respective ability level, as determined by MAP specifications. Responses to each pair of complementary items were then combined before dichotomizing.

In addition, the explanatory analysis includes a pair of dichotomous covariates theorized to influence the DIF fidelity: grade level (fourth, fifth) and experimental condition (treatment, control). I used dummy coding (0, 1) for the DIF analysis and effects coding (1, -1) for the actual IRT and ANOVA analyses.

Analysis plan

The analysis of survey data progresses in three stages: checking IRT and ANOVA assumptions for the current data, establishing the proper explanatory IRT analysis by comparing data fit between different models, and both ANOVA and IRT analyses to explain responses using covariates.

Assessing IRT assumptions

It is necessary to assess each item for DIF with respect to each covariate. DIF is a violation of the IRT assumption of unidimensionality and manifests as biased scores that are higher for one group than the other even after accounting for differences in the trait. Any reasons given for suspecting DIF would be purely speculative (or informed guessing at best), and would be more appropriate for a separate study with that focus. For the DIF analysis, treatment and fourth grade are considered focal groups (coded 1), while control and fifth grade are reference groups (coded 0).

To assess dimensionality with respect to relevant subgroups, I use both observed conditional invariance (OCI) and unobserved conditional invariance (UCI) DIF detection procedures (Millsap and Everson, 1993). OCI procedures use total observed scores to compare groups' item responses, while UCI procedures make latent trait comparisons. There is no single "best" method for assessing DIF, and so it is wise to use a number of methods and suspect DIF if it is detected by several distinct methods. Each of the approaches has relative strengths and weaknesses, and multiple detection indicates the likelihood that true DIF is present. As such, I employ several methods[6]: three OCI procedures, the Mantel-Haenszel statistic (MH, Mantel &

[6]A larger number of DIF detection procedures was applied here than typically would be; in typical practice, a researcher may test for DIF using as few as one OCI and one UCI method. Using ten methods ensures a very high

Haenszel, 1959), the standardized approach (Dorans & Kulick, 1983) and the logistic regression (LR, Swaminathan & Rogers, 1990) model; and the UCI likelihood ratio procedure (Thissen, Steinberg, & Wainer, 1988; Thissen, Steinberg, & Gerrard, 1986).

The MH procedure is used to test whether the odds of a correct response for the focal and reference groups are independent conditional on the observed score. With the standardized approach, one evaluates the weighted difference of proportion correct between the focal and reference groups. The LR procedure allows one to utilize model fit so as to determine whether responses to an item are best explained by a person's total score alone, by also accounting for group membership (uniform DIF), or by further accounting for an interaction between total score and group membership (non-uniform DIF). I use three LR methods that check separately for uniform DIF, non-uniform DIF, and both. With the likelihood ratio procedure, one compares fit when item parameters are constrained to be equal between groups or allowed to vary. All of these analyses are conducted with syntax from several open-source program libraries (the Matrix, lattice, lme4, ltm, and difR packages) in R (R Development Core Team, 2012). For the R code used, see Appendix B.

Applying these DIF methods collectively yields ten tests of DIF per item: a single effect size is produced using the MH method; with the standardized and LR methods, one obtains multiple types of effect sizes that represent different approaches to testing DIF (i.e., DSB and ETS for standardized, ZG and JT for each of the three LR tests)[7], and the likelihood ratio test in the R lmer library produces only a *p*-value. Because the teacher survey is designed to measure

---

standard for determining DIF, important since DIF items will be omitted entirely from the analysis, and also has the benefit of providing examples for interpreting a wide array of DIF indices.

[7] Dorans, Schmitt, & Bleistein's (2002) proportion of correct responses by group and total score (p-DIF) and the transformation to the Educational Testing Service's delta scale (Dorans, 1989); and the change in R^2 between models on the respective Zumbo & Thomas (1997) and Jodoin & Gierl (2001) scales.

the three dimensions of DI through three separate subscales, unidimensionality with respect to grade and experimental condition must be assessed separately within each subscale. My a priori process for identifying DIF items is to suspect an item of DIF if at least half of the effect sizes[8] indicate moderate or large DIF. I then inspect the suspected items in an exploratory manner to identify any similarities among items that may suggest a cause for the DIF. A confirmatory analysis for DIF would allow for testing any resulting hypotheses, but that is beyond the scope of this study.

When DIF is detected during scale development, items should be revised or replaced and the instrument re-piloted. This is not possible when data already have been collected, so responses to DIF items must be excluded from the analysis. One approach to comparing CTT and IRT would be to exclude DIF items only from the data used for IRT analysis; however, that approach here would put undue emphasis on the data cleaning process and make it impossible to attribute differences in results specifically to the analytic methods. Therefore, I will remove any DIF items from both analyses, leaving only the relatively unidimensional items.[9]

Finally, I will examine item-level fit statistics (mean squared and *t*-values) to determine whether an item fits to the data well.

ANOVA model and assumptions

I will use a two-way between-groups ANOVA (Kirk, 1968) with raw scores as a CTT approach comparable to the IRT procedure. The assumptions of this ANOVA model are:

[8] There is no agreed-upon standard for the number of DIF tests one should run or how many must be significant to suspect DIF (although using multiple methods compensates for weaknesses of individual methods). A sensitivity test conducted after the main DIF analysis indicated that results would have been unchanged had I increased or decreased the cut-off by one (i.e., omitting items with four or six significant effect sizes).

[9] Unidimensionality is not an assumption of CTT methods in general or ANOVA in particular, but eliminating DIF items renders the interpretation of CTT results in the context of a single target construct more valid (Reise & Henson, 2003).

1. Independence of observations,

2. Normally distributed residuals, and

3. Equality of variance (homoscedasticity) across groups.

I assess independence by examining the research design to identify likely sources of dependence and calculate the intraclass correlation (ICC) as the proportion of total variance in responses that is due to clusters (i.e., school districts) with equation 3 (Cohen, et al, 2003):

$$ICC = \frac{MS_{Tx} - MS_e}{MS_{Tx} + (\tilde{n} - 1)MS_e}, \quad \tilde{n} = M_n - \frac{sd^2(n_j)}{gM_n} \tag{3}$$

where,

- MS_{Tx} indicates the mean squared treatment,
- MS_e indicates the mean squared error,
- ñ indicates the adjustment for unequal groups,
- M_n indicates the mean group size,
- $sd^2(n_j)$ indicates the variance of the group sizes, and
- g indicates the number of groups.

Normality is assessed separately for each level of grade level and experimental condition, as well as for each grade by condition combination. To do so, I visually inspect distributions in histograms, Q-Q plots, and box plots, I measure deviations from the normal curve skew and

kurtosis, and I test the significance of non-normality with the Shapiro-Wilk test (Shapiro & Wilk, 1965). I use Levene's test of the absolute value of the deviation from the mean (Levene, 1960) to assess homoscedasticity among the between factors. All ANOVA assumption checking is conducted in IBM SPSS 20.0 (2011).

Statistical adjustments and alternate (e.g., multivariate) tests can be employed to compensate for some ANOVA assumption violations. For example, transformations and Winsorizing outliers can improve non-normality and heteroscedasticity. However, some adjustments are not feasible in this study: altering the distribution of responses could obscure the defining characteristics of high and low fidelity data.

Descriptive IRT analyses

Once DIF items are detected and accounted for, I test three separate models for the DI constructs using the R lmer function (Bates & Maechler, 2009; see Appendix B for R code used). The lmer function in the ltm4 R library uses the Laplace approximation to the integrand to approximate the maximum likelihood estimation (MLE) of model parameters, and it is relatively efficient compared to several other methods (Cho, Athay, & Preacher, 2012).

First, I test the model while ignoring multidimensionality as in equation 4:

$$\text{logit}\left[P\left(y_{ji} = 1|\theta_j\right)\right] = \left(\theta_j - \beta_i\right) \quad (4)$$

where

- j indicates a person ($j = 1, \ldots, J$),

- i indicates an item ($i = 1, \ldots, I$),
- β_i indicates the fixed item difficulty parameter, and
- θ_j indicates the trait parameter, random across persons, assumed to be approximately normally distributed, $\theta_j \sim N(\mu, \sigma)$.

Next, I test the model by estimating parameters in three separate, uncorrelated models (for each dimension) of the form of equation 5:

$$\text{logit}\left[P\left(y_{jid} = 1 \middle| \theta_{jd}\right)\right] = \left(\theta_{jd} - \beta_{id}\right) \quad (5)$$

where

- d indicates a dimension ($d = 1, 2, 3$) of DI,
- β_{id} indicates the fixed item difficulty parameter for each item in a given dimension, and
- θ_{jd} indicates the trait parameter for each person, random across persons, assumed to be approximately normally distributed, $\theta_{jd} \sim N(\mu_d, \sigma_d)$.

The descriptive MIRT model that I will employ is derived from the generalized descriptive longitudinal item response model (Cho, Athay, & Preacher, in press), but without the longitudinal component, as subconstructs of DI rather than time points are nested in persons. The generalized descriptive model, which excludes person group covariates to explain person responses, is presented in equation 6:

$$\text{logit}[P(y_{jdi} = 1|\sum_{d=1}^{D}\theta_{jd}]) = (\sum_{d=1}^{D}\theta_{jd}) - (\sum_{d=1}^{D}\beta_{id}) \quad (6)$$

where

- d indicates the dimensions (d = 1, 2, 3) of DI,
- β_{id} indicates the fixed item difficulty parameter for each item in a given dimension, and
- θ_{jd} indicates the trait parameter for each dimension, random across persons, assumed to be approximately normally distributed, $\theta_{jd} \sim N(\mu_d, \sigma_d)$.

This type of MIRT model is called a compensatory model (Gelman and Hill, 2007; Henson, Templin, and Willse, 2009), because a decrease in one dimension can be compensated for in the overall score by an increase in another dimension. In context, this means for example that a teacher constrained in their instructional strategies (DS) can compensate by using more diverse content (IT) to improve overall DI.

Each of these models includes the item difficulty parameters (β) but not parameters for discrimination or guessing, potentially capturing less information than either 2PL or 3PL models. However, there is some evidence from simulations (DeMars, 2001) that little is lost using the 1PL model when the primary interest is group differences.

The three models will be compared using model fit statistics, specifically AIC (Akaike, 1974), BIC (Schwarz, 1978), and deviance statistics calculated from log likelihood. Although

model specification for the explanatory analysis is driven primarily by the theoretical composition of the DI general construct, a superior fit to the descriptive MIRT model would provide support that the model has not been misspecified.

Explanatory IRT and CTT analyses

The generalized explanatory model (GEM) differs from the descriptive model mainly in that it includes the grade level and experimental condition covariates thought to explain variations in DI implementation. The GEM parameter estimates can be considered in four parts: 1) the dimensions of DI as within-person random effects; 2) the person group (grade level, experimental condition, and their interaction) as between-person fixed effects; 3) the fixed effects the dimension interactions with each between-effects factor; and 4) the item random effects. I will not interpret the within and within-between effects for two reasons. First, unlike the strictly between-groups effects, the non-linear, interval nature of the IRT scale precludes directly comparing within-person effects with their CTT equivalents. Second, the objective of this analysis is to compare CTT and IRT methods for assessing group differences across all dimensions, not how groups differ from one dimension to another, which would be analogous to a split-plot ANOVA rather than a strictly between-subjects ANOVA.[10]

Instead, I interpret only the between-subjects fixed effects across the single, integrated DI dimension. This is possible because the GEM aggregates the common variance among the IM, DS and IT dimensions into a single factor via their inter-correlations. The interpretation of these results can be compared to the main and interaction effects in a between-groups ANOVA, where

[10] This approach to creating a composite dependent variable for teachers' self-reported intervention fidelity is the same as was employed for the original CTT analysis of MAP data (Cordray et al, 2011).

individuals' scores for each of the three dimensions have been averaged together into a composite dependent variable.

The ANOVA gives equal weight to each level, even if there is no theoretical or empirical justification for doing so (as is the case for the subconstructs of DI). In contrast, the GEM is able to estimate the proportional contribution of each subconstruct to the latent DI dimension, guided by fitting empirical data to a theoretically supported model.

The GEM is described by equation 7:

$$\text{logit}[P(y_{jcdig} = 1| \sum_{d=1}^{D} \theta_{jd}, \epsilon_i)] = \zeta_{..} + \zeta_{c.} + \zeta_{.g} + \zeta_{cg} + (\sum_{d=1}^{D} \theta_{jd}) - \left(\sum_{d=1}^{D} \beta_d + \epsilon_i\right) \quad (7)$$

where

- c indicates the experimental condition covariate,
- g indicates the grade level covariate,
- $\zeta_{..}$ indicates a fixed intercept parameter (average score across persons and items),
- ζ_{cg} indicates a fixed interaction parameter of condition and grade level,
- $\zeta_{c.}$ indicates a fixed parameter of condition alone,

- $\zeta_{.g}$ indicates a fixed parameter of grade,
- $_{\beta d}$ is a fixed item difficulty parameter,
- ϵ_i is an item difficulty parameter, random across items and assumed to follow an approximately normal distribution $N(0_{3x1},\sigma^2_{\varepsilon})$, and
- θ_{jd} is a random person trait parameter for each dimension d (average ability across persons), and is assumed to follow an approximately multivariate normal distribution, with $\boldsymbol{\theta}_{jd} = [\theta_{j1}, \theta_{j2}, \theta_{j3}]' \sim MN(0, \boldsymbol{\Sigma}_{(3x3)})$

The random item parameter ϵ_i is interpreted in this context as item-specific deviance from the mean difficulty of the dimension-specific item groups, where σ^2_{ε} is a random residual (De Boeck & Wilson, 2004). The chief benefit here of random item effects in the model is that, because it includes a single item variance parameter rather than separate parameters for each item, the number of parameters to be estimated using the lmer package in R are reduced. The GEM analysis here produces between-group results that can be interpreted similarly to the results of a 2 (grade) x 2 (condition) between-groups ANOVA.

I use the results of this analysis to address two questions related to DI implementation:

1. Main effects: Controlling for other factors, how does fidelity differ between: a) grade levels and b) experimental conditions?

2. Two-way interaction: Controlling for other factors, how does fidelity differ between grades by experimental conditions?

As the CTT comparison, I will also use a between-groups ANOVA design, with the general model presented in equation 8:

$$Y_{mn} = \mu + \alpha_m + \beta_n + (\alpha\beta)_{mn} + \epsilon_{mn} \quad (8)$$

where

- Y_{mn} indicates the individual person test score
- μ indicates the grand mean for the population,
- α_m indicates effects at levels of the first between factor,
- β_n indicates effects at levels of the second between factor,
- $(\alpha\beta)_{mn}$ indicates the between groups interaction effect, and
- ϵ_{mn} is the error term (population model only).

The ANOVA has been shown in simulations to have other potential weaknesses compared to IRT models due to not achieving interval-scale measurement. In different simulation studies, Embretson demonstrated how ANOVA's translation of interval data into ordinal measures can create spurious interactions or fail to detect real ones (Embretson, 1996), and can bias the results of between-groups comparisons (though apparently only on the order of 0.025 to 0.05 standard deviations, based on graphed results) (Embretson, 1994).

Results

IRT assumption checking

I used OCI and UCI DIF detection methods for each dimension of DI to assess whether the survey items met the assumption of unidimensionality, thus being invariant across grade levels and experimental conditions. Of the 31 items across three subscales, 8 were suspected of DIF: none of the 5 IM items; among the 19 DS items, 3 for condition, 3 for grade, and 2 for both; and none of the 6 IT items.

Table 1. DIF items with respect to grade or condition

The common stem for all DS scale items reads: "Looking back over the school year and thinking about how you taught reading/language arts to these high (or low) -achieving students, how often did you…"

Item	MH ES[1]	Standard DSB[2]/ETS[1]	LR Both ZT[3]/JG[4]	LR Uniform ZT[3]/JG[4]	LR Non-U, ZT[3]/JG[4]	Likelihood Ratio, p[5]
Items with DIF by grade level:						
9. Assign reports	**C**	A/**C**	A/**C**	A/**C**	A/**B**	.67
10. Assign projects or other work requiring extended time for students to complete	**C**	**C**/**C**	A/**B**	A/**B**	A/A	.29
11. Make time available for students to pursue self-selected interests	A	**B**/A	**B**/**C**	A/A	**B**/**C**	.45
17. Use enrichment centers	**C**	**C**/**C**	**B**/**C**	A/**C**	A/**B**	**< .01*** **
24. Encourage student participation in discussions	**C**	**B**/**C**	A/**B**	A/A	A/**B**	.26
Items with DIF by experimental condition:						
6. Use basic skills worksheets	**C**	**C**/**C**	**B**/**C**	**B**/**C**	A/A	**< .01*** **
7. Use enrichment worksheets	**C**	**C**/**C**	**B**/**C**	A/**C**	A/A	**.01***
9. Assign reports	**C**	A/**C**	A/**C**	A/**C**	A/**B**	.67
11. Make time available for students to pursue self-selected interests	A	**B**/A	**B**/**C**	A/A	**B**/**C**	.45
23. Use computers	**C**	**C**/**B**	A/**B**	A/**B**	A/A	**.04***

Items in shaded rows are biased toward the comparison group (fifth grade or control), while items in non-shaded rows are biased toward the focal group (fourth grade or treatment).
Medium and large effect sizes and significant *p*-values are bolded.
1. ETS delta effect size codes: A = 0.0 to 1.0, B = 1.0 to 1.5, C = 1.5+
2. Dorans, Schmitt & Bleistein effect size codes: A = 0.00 to 0.05, B = 0.05 to 1.00, C = 1.00+
3. Zumbo & Thomas effect size codes: A = 0.00 to 0.13, B = 0.13 to 0.26, C = 0.26 to 1.00
4. Jodoign & Gierl effect size codes: A = 0.00 to 0.035, B = 0.035 to 0.07, C = 0.07 to 1.00
5. *P*-values (alpha = 0.05): * = .01 to .05, *** < .01

It appears that differentiating instructional strategies (DS) may be a more difficult construct to measure unidimensionally by self-report than content or grouping. Once these 8 items were confirmed as multidimensional, I investigated the consequences of removing the items for the subscales using ACER ConQuest 2.0 (Wu, 1997). First, I examined each scale's Wright map (see Appendix D), a chart that displays item difficulties and individual trait levels on a common logit scale (Wilson, 2003). If the difficulties of the removed items align with the trait levels most of interest (i.e., high fidelity) and no other items assess those levels, then standard errors for scores will increase where precision is most needed (Hambleton & Jones, 1993). Most of the DS items' difficulties are in ranges covered by other items as well, but items 4 and 19 are alone at the extreme positive and negative ranges of the trait, respectively. As a result, removing the 8 items reduces the test information curve (TIC, Figure 3). Information in this context is the inverse of the standard errors, and so the precision for measuring trait levels where items were removed has been decreased as well.

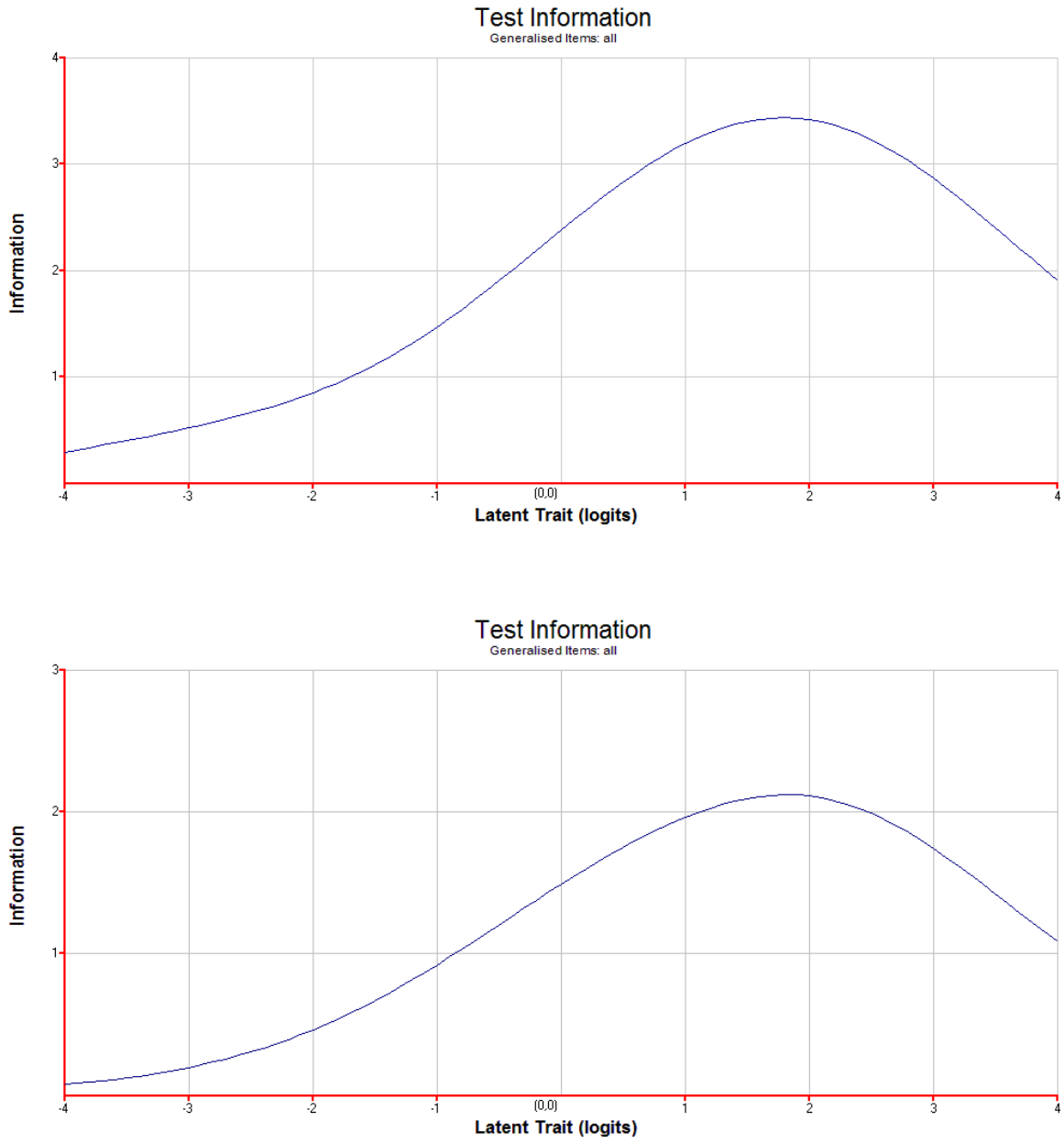


Figure 3. DS test information curve before removing DIF items (top) and after (bottom). Note that the information scale on the x-axis in the top graph (range 0 to 4) differs from the information scale in the bottom graph (range 0 to 3).

Comparing the TICs, it can be seen that information decreases for individuals across the trait spectrum, by two-fifths up on the lower range up to about a half at extremes of the scale.

I also looked at the subscales' separation reliability (SR), the proportion of variance in responses explained by the model, both before and after removing DIF items:

Table 2. Impact of omitting DIF items on DI subscales' separation reliability

Dimension	SR with DIF items	SR without DIF items
IM	.97	.97
IT	N/A	.97
DS	.98	.97

The separation reliabilities for both IM and DS decreased only slightly (each by .01 or less) and the value for all scales is near 1.00, indicating that nearly all of the variance in responses is explained by the model as a whole.

Similarly, I calculated the CTT reliability statistic Chronbach's alpha (Chronbach, 1951), the internal consistency of the scales, with and without DIF items, as well as their standard errors of measurement (SEM), the individual standard errors' mean (in Table 3).

Table 3. Impact of omitting DIF items on the DI scale's and subscales' internal consistency

Dimension	α with DIF items	α without DIF items	SEM with DIF items	SEM without DIF items
IM	N/A	.38	N/A	0.85
IT	N/A	.75	N/A	1.04
DS	.80	.76	1.56	1.21

Removing DIF items from the DS subscale marginally decreased the internal consistency and increased the SEM for each, indicating greater precision.

Removing the DIF items improved some scales' CTT reliabilities but resulted in less IRT information and precision for some ranges of DI. Weighing these impacts, I concluded it was necessary to omit the DIF items to avoid violating the IRT assumption of unidimensionality.

Finally, I examined the IRT fit statistics for each item in the revised subscales (see Appendix E for all item statistics). Items with good fit should have weighted means square fit statistics (MNSQ) between 0.75 and 1.33 and *t*-values between -4 and 4. Standard errors are higher for items at extreme levels of difficulty, and therefore the accuracy of the estimate and amount of information provided are lower. Across all three subscales, the MNSQ fit ranges from 0.78 to 1.14, and the absolute value of all *t*-values never exceeds 1.50. As item fit statistics are all within acceptable ranges for all three subscales, I conclude that the fit is adequate.

ANOVA assumption checking.

The between-groups ANOVA relies on three assumptions: 1. independence, 2. normality, and 3. homoscedasticity. Independence at the school level was achieved in the MAP study

sample by virtue of the CRT design: grade levels were randomly assigned to condition within school, essentially matching experimental teachers with control teachers to control for clustering effects.[11] At the district level, the intraclass correlation (*ICC*) was calculated as (0.034 - 0.035)/[0.034 + (29.49 - 1)*0.035] = -0.0008, with an adjustment for unequal groups of $\tilde{n}$ = 33.6 – [690.3/(5*33.6)] = 29.49. Such small, negative *ICC*s indicate approximately zero dependence due to clusters.

Levene's Test yielded a non-significant result (F = 0.47 [3, 164], p = .70), showing no evidence of violations of homoscedasticity. The cell *N*, skew, kurtosis, and Shapiro-Wilk statistics (null hypothesis of normality) are presented in Table 4.

Table 4. Assessing the ANOVA normality assumption

	Grade 4	Grade 5	Treat.	Control	G4 x Tx	G4 x C	G5 x Tx	G5 x C
N	86	82	83	85	52	34	31	51
Skew	-0.07	-0.66	-0.34	-0.32	0.17	-0.29	-1.18	-0.43
Kurtosis	-0.26	0.07	0.45	-0.15	-0.58	-0.02	1.92	-0.27
S-W	0.99	0.95*	0.98	0.98	0.98	0.98	0.91*	0.96

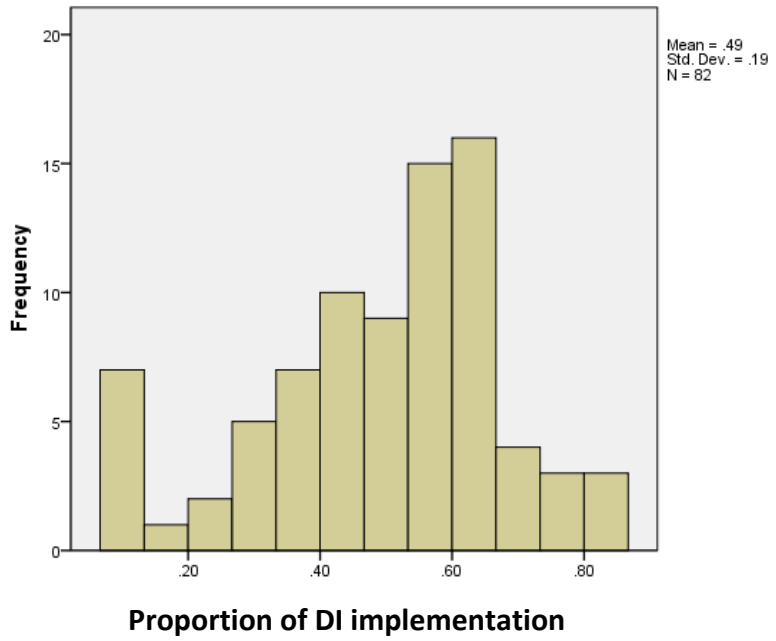
*p < .50
Note: Kurtosis values are the amount of deviation from the normal distribution value of 3

Though no main factor levels demonstrated unacceptable values for skew (absolute values greater than 1) or kurtosis (absolute values greater than 2) (Harlow, 2005), the distribution

[11] As noted by Cordray et al. (2011): "Because there are relatively few teachers in each group and each grade, we assume that the ICC is 0.00."

for grade 5 was found to have a significant Shapiro-Wilk statistic. At the cell level, the distribution of grade 5 x treatment condition responses also departed significantly from normality, as assessed by Shapiro-Wilk, and was above acceptable levels for skewness.

Inspection of the graphs confirmed only slight negative skewness for fifth grade responses (see on next page Figure 4, top) but major departures from normality for fifth grade treatment teachers' responses: the histogram confirmed that the distribution is highly negatively skewed and moderately leptokurtic (Figure 4, bottom), the Q-Q plot deviated from the diagonal, and the box plot showed outliers at the lower end of DI scores. Skewness is typically a minor concern for the generally robust ANOVA (when group sizes are equal or nearly equal; Field & Miles, 2010).



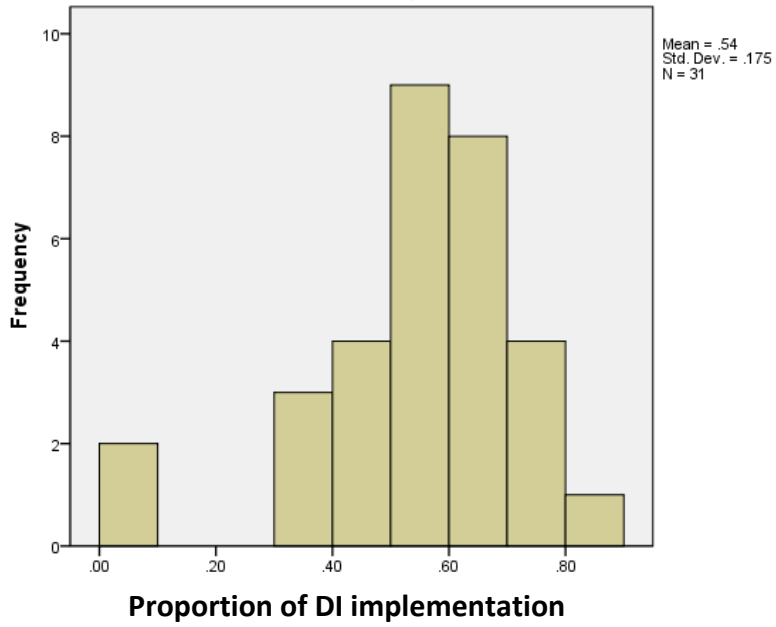


Figure 4. Distributions of responses from all fifth grade teachers (top) and fifth grade treatment teachers (bottom)

Options for addressing the non-normality include transforming the data, Winsorizing outliers, and selecting an alternative non-parametric test. I chose not to adjust the data to achieve normality because the deviations may not be mere statistical noise: they are consistent with theory, in that they may result from a large proportion of teachers implementing DI at a relatively high level while a much smaller proportion implemented at varying low levels. I did not select a non-parametric method because, unlike ANOVA, non-parametric alternatives to the between-groups ANOVA share IRT's disadvantages of being unfamiliar to many researchers and, in cases relevant to this particular design, are not included in many widely-accessible software packages (Maxwell & Delaney, 2004).

Descriptive IRT model results

I analyzed teachers' responses to the revised scales using a sequence of IRT models with the R lmer package, allowing me to compare model fit to the data and select the model that best describes the DI construct and subconstructs. The three models I tested were a single unidimensional model, consecutive unidimensional models for each dimension, and a multidimensional model allowing correlations between the three dimensions. The fit of these models as assessed by AIC, BIC, and deviance statistics, given in Table 5.

Table 5. Descriptive IRT Model Fit Comparisons

Model	AIC	BIC	Deviance
1-Dimensional model	3855.0	4006.0	3807.0
Consecutive models	3853.2	3987.6	3801.2
3-Dimensional model	3784.0	3966.0	3726.0

The consecutive models showed no improvement in fit over the unidimensional model, with AIC and BIC each minimally reduced, and a non-significant chi-square difference: 3807 (24) – 3801.2 (28) = 5.8 (4), $p = .22$. Compared to the unidimensional model, the GEM has smaller AIC and BIC values, indicating improved fit. The deviance difference chi square statistic was consistent with improvement, rejecting the null hypothesis of no difference: 3726 (24) – 3632 (29) = 94 (5), $p < .01$.

The fit of the GEM and its consistency with the theory of the DI construct in MAP validate use of the model for the explanatory analysis. The survey responses will be fitted to the three-dimensional logistic Rasch model with correlated dimensions, formally equivalent to the explanatory general longitudinal item response model (Cho, Athay & Preacher, in press).

Explanatory IRT and ANOVA results

Though I am not interpreting within-group factors from this analysis, I will note that the inter-correlations among dimensions are increased (IM and DS = 0.96, IM and IT = 0.92, DS and IT = 0.82) from the descriptive model correlations (IM and DS = 0.78, IM and IT = 0.65, DS and IT = 0.64). These results are consistent with the prediction that grade level and experimental condition are related to the general DI implementation through its subconstructs. Model fit and

estimates of fixed effects are presented in Tables 6 and 7, respectively. Item fit indices for this model (see Appendix E) are all within acceptable ranges, and the Wright map (see Appendix D) shows that items are well-dispersed across the entire DI trait distribution, with more located at higher levels of fidelity.

Table 6. Explanatory GEM fit

AIC	BIC	logLik	Deviance
3844	3963	-1903	3806

Table 7. Explanatory GEM estimates of fixed effects

	Estimate	Std. Error	Odds Ratio	*z*	Pr(>|*z*|)
Grade	0.45	0.30	1.56	1.48	.14
Condition	0.20	0.30	1.23	0.67	.50
Grade x Condition	0.27	0.44	1.30	0.61	.54

The estimates in Table 8 should be interpreted as the logit difference between levels of grade and condition, so that differentiation of instruction was 0.45 logits greater among fourth grade teachers than fifth grade teachers, and 0.20 logits greater among teachers in the treatment compared to the control condition.

None of the between-groups main or interaction effects were significant, with the smallest *p*-value being for the comparison of grades. Based only on the explanatory GEM analysis, a researcher should fail to reject the relevant null hypotheses and conclude that there is no evidence that DI implementation varied between grades or treatments or in a significant interaction.

Looking only at the between-groups results of the between-groups ANOVA, presented in Table 8, would lead to a different conclusion.

Table 8. ANOVA between-groups effects

Source	Type III SS	df	Mean Square	*F*
Intercept	134.82	1	134.82	1319.40
Grade	0.42	1	0.42	4.09*
Condition	0.23	1	0.23	2.23
Grade x Condition	0.14	1	0.14	1.38
Error	16.76	164	0.10	

*Significant at the $\alpha = 0.05$ level

The ANOVA found a significant difference between grade levels ($p = .05$), and so the null hypothesis of no difference should be rejected. As such, a researcher should conclude that there is evidence that DI implementation differs between grade levels but not for condition or their interaction. The achieved relative strength (ARS, Hulleman & Cordray, 2009) $g = 0.38$ is relatively small considering that the effects of this cause would likely be of smaller magnitude.

The conflicting conclusions based on different methods confirm that the choice between IRT and CTT approaches for analyzing group differences in fidelity can be consequential, and imply the possibility that IRT methods may be less problematic. Technically, the two sets of *results* are not necessarily contradictory, because the CTT and IRT methods test different null hypotheses about group true scores and latent trait levels, respectively. In practice, however, the *interpretations* of CTT results generally imply that we have contrasted the same sort of underlying trait with scale invariance as in IRT. Even setting aside that issue, the MIRT model would still be considered preferable if it reached the correct conclusion because it was able to make use of greater information about each item and from correlations among dimensions.

This outcome is not surprising despite related studies finding few meaningful differences: IRT has many theoretical advantages that are especially likely to apply when analyzing fidelity data, and *p*-values make for a more straightforward comparison across theories than do parameter estimates or effect sizes. Granted, the present methodology does not clarify which approach led to a more accurate assessment of the difference between grades or why, but there are some clues to be found in the descriptions of the two procedures and their assumptions

Conclusions

In this study, I have compared the assumption-checking procedures, models, and results of IRT and ANOVA analyses using particular models and an empirical dataset for assessing intervention fidelity. Interestingly, the two approaches led to conflicting statistical conclusions (GEM rejects but ANOVA fails to reject the null hypothesis) about differences in intervention fidelity between grade levels. To explore why this might be so, I revisit the four aspects of

intervention fidelity assessment listed in the introduction as potentially benefiting from differences between IRT and CTT approaches, as well as an additional item that is more general.

1. Latent constructs versus manifest variables. In addition to matching the construct-level understanding of DI with a construct-level analysis of group ability levels, the MIRT model also estimated item difficulties relative to the latent trait scale. The amount of information an item provides (and the uncertainty of that estimate) was taken into account, whereas the ANOVA treated all items as equally precise and equally difficult.

2. Multidimensionality. By assessing latent constructs, the IRT DIF procedures were able to detect items that were biased toward one group over another; however, these items were removed from the data analyzed by both groups and so could not have influenced differences in results.

3. Other item and scale characteristics. The distribution of responses with respect to grade, as described under CTT assumption checking, were somewhat non-normal and violated an assumption of the ANOVA procedure. Without being able to rule out a substantive cause (as would be common for fidelity), it was deemed unwise to adjust the data. IRT does not depend on this assumption, and its interval scale is less likely to distort response distributions. However, violating normality can inflate ANOVA type I error rates (Maxwell & Delaney, 2004; Field & Miles, 2010), in which case the non-normality could be the cause of erroneously finding an effect for grade. There is no accepted procedure for calculating power for the group difference in using an IRT analysis, so it remains possible that the failure to detect a difference between grades is an error.

4. Sample size. An IRT analysis can fail to detect differences between groups if the sample size is too small and the model fits poorly, and there is not yet a definitive way to assess the sample size design for power of the group difference detection to achieve the specified the effect sizes in using an IRT analysis, particularly one with multiple dimensions and subgroups, and with non-normal distributions for each dimension-by-subgroup scores. Item fit statistics indicated good fit for all items, and the multidimensional model was found to have a superior fit to the other models tested.

5. Interval scale. The benefits of interval scaling are not specific to fidelity analyses and so were only briefly discussed in the Introduction and Methods sections. Reducing interval-level traits to ordinal-level responses will result in misleading conclusions if the data are interpreted as though equal differences in observed scores represent equal differences in ability at all points along the scale. Embretson (1994) showed in simulations that ANOVA can produce biased results because it fails to achieve interval-level measurement. The simulations only revealed small differences in effect sizes (by about 0.025 to 0.05), but the effect of grade level was only marginally significant ($p = 0.045$) and would have been in agreement with the GEM non-significant results with only a slightly smaller effect.

Discussion

Perhaps the most important contribution this study could make would be to demonstrate for those unfamiliar with IRT the GEM model and its advantages when compared with traditional approaches. Indeed, it is likely that many researchers do not consider IRT analytic approaches for RCT data (implementation or impact) because they perceive IRT to be too

complicated, because they do not understand the theoretical rationale, or because they believe IRT models can only be applied when there are thousands of responses or hundreds of items. The GEM model can be grasped more easily in comparison with the ANOVA model, and the requisite procedures are hardly more onerous given their benefits. The GEM model is also appropriate for the relatively small sample sizes in most education RCTs, as well as for fidelity measures with relatively short scales.

The results of the IRT and ANOVA analyses clearly differ regarding the significance of the grade level factor for DI fidelity. Because these are empirical outcomes, it is not possible to discern the true population means at each grade level and assess which conclusion is correct. Nor do the statistical results reveal precisely why the two models differ, or give any indication of which model's assumptions were less tenable or whether manifest or latent DI was a better approximation of reality. Yet the results do suggest that researchers choosing between IRT and CTT methods to analyze intervention fidelity data are a priori including in their design an unaccounted for factor, unrelated to their research questions, that may nonetheless influence how their results are interpreted and applied.

Any or none of the factors listed in the conclusions section may have contributed to the difference in results. Importantly, that means we cannot yet predict the particular IRT and CTT tests that will yield different results, under what conditions and in what way, preventing researchers from selecting the optimal test of intervention fidelity for a planned study. Though it is not possible to know which result is more correct, advocates of IRT might point to simulation studies as evidence that conclusions consistent with IRT should be trust over those consistent with CTT when they disagree. Until there is conclusive evidence favoring IRT over CTT (which

this study certainly does not provide), most researchers outside of large-scale ability testing likely will continue to employ IRT methods sparingly, if at all.

In the meantime, researchers should recognize the potential limitations of CTT when interpreting classical results. For example, they can acknowledge their results from statistical tests performed on true scores may vary between scales, just as they currently must justify generalizing results to a different population. This may be particularly relevant in the context of fidelity assessment, where instrument design is determined by conceptions of constructs that may vary widely; and in education research generally, where intervention effects are far more likely to be detected by research-designed instruments than by the standardized tests that increasingly determine school and teacher evaluations. Of course, researchers would be ill-advised to employ the unethical and statistically problematic practice of using both CTT and IRT methods and then simply report the more favorable results.

Some might argue that the convenience of CTT methods, being widely taught and available in user-friendly software, trumps any advantages that may or may not be provided by the more esoteric methods of IRT. A similar argument has been made that non-parametric methods have little value except in cases of extreme assumption violations, and the retort is the same for both: a researcher ought to choose the method that is most powerful when meeting acceptable levels of precision (Maxwell & Delaney, 2004). And while those only casually acquainted with IRT methodology and interpretation may have difficulty proceeding through IRT assumption tests, model fit comparisons, and explanatory analyses, these procedures are no more inherently cryptic than for many CTT procedures. Furthermore, the software used for most of the present analysis (R Development Core Team, 2012) can be downloaded from the Internet for free.

A major weakness of this study is that results for this dataset cannot be generalized: neither the data nor the particular IRT and CTT models were selected at random. Instead, this study serves as a proof by counterexample, in that finding inconsistent results for one study proves that IRT and raw score analyses will sometimes lead to different conclusions. Another shortcoming of the empirical approach noted repeatedly above is that showing that the two statistical conclusions contradict does not reveal which conclusion is right, if either, or why the other is wrong.

Another limitation on generalizability of the study's conclusions is that the overall analytic plan is somewhat unrealistic. One would never conduct multiple tests of significance for the same null hypothesis, and one would also not typically conduct an IRT DIF analysis before using raw scores (although it could benefit from doing so). A comparison truer to the real world would be between the scale as altered and analyzed by IRT methods, and the intact scale as analyzed by CTT methods. Unfortunately, the results would not reveal whether differences in results were due entirely to the models or should be attributed to the different items. Alternately, the comparison might be made for an IRT-developed scale, although it is also uncommon for such a scale to be analyzed using CTT methods.

Similarly, the generalization of this study's conclusions to other CTT methods may be limited. Certainly, theory indicates that the so-called shortcomings of CTT should apply to all methods, but the purpose of an empirical study is to confirm theory. For example, it would be fair for those skeptical of IRT methods' alleged superiority to insist that the study be replicated with factor analysis and structural equation modeling with raw scores.

Finally, both the IRT and ANOVA analyses involved decisions in cases where there are no absolute decision rules. When assessing DIF, researchers have used numerous strategies, from applying a single DIF test to multiple, with varying cut-off points. To be conservative in judging multidimensionality, I chose to use six tests with ten effect sizes, requiring a minimum five significant effect sizes to suspect DIF for an item. There is also no universal standard for how much non-normality the ANOVA can tolerate, nor is there a single remedy or alternative. I chose not to adjust the data to achieve non-normality for substantive reasons, but some might consider this to have been an unfair test of the ANOVA, especially given that I eliminated items to ensure unidimensionality for IRT.

REFERENCES

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Akaike, A. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Amin, L. (2011). *Psychometric Methods to Develop and to Analyze Clinical Measures: A Comparison and Contrast of Rasch Analysis and Classical Test Theory Analysis of the PedsQL 4.0 Generic Core Scales (Parentreport) in a Childhood Cancer Sample*. Retrieved from Digital Commons at McMaster.

Baranowski, T., Allen, D.D., Masse, L.C., & Wilson, M. (2006). Does participation in an intervention affect responses on self-report questionnaires? *Health Education Research*, *21*, 98–109.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed–effects models using s4 classes. Retrieved August 14, 2012 from: http://cran.r-project.org/web/packages/lme4/index.html.

Blank, R. K., Porter, A., & Smithson, S. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science.* Report from Survey of Enacted Curriculum Project (National Science Foundation REC98–03080). Washington, DC: Council of Chief State School Officers.

Boomsma, A., & Hoogland, J. J. (2001). The Robustness of LISREL Modeling Revisited. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 139–168). Lincolnwood, IL: Scientific Software International.

Chen, H.T. (1990). *Theory-driven evaluation.* Thousand Oaks, CA: Sage Publications.

Cho, S.J., Athay, M., & Preacher, K. J. (in press). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology*.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences, third edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Cordray, D.S., Pion, G.M., Dawson, M., and Brandt, C. (2008). *The Efficacy of NWEA's MAP Program. Institute of Education Sciences funded proposal.* Unpublished Technical Report to Institute of Education Studies (IES) under contract ED-06-CO-0019.

Cordray, D.S., Pion, G., Dawson, M, and Brandt, C. (2010). *The Effects of the Measures of*

Academic Progress (MAP) Program and its Associated Training on Differentiated Instruction and Student Achievement: Analytic Plan, Selected Revisions. Unpublished Technical Report to Institute of Education Studies under contract ED-06-CO-0019.
.

Cordray, D. S., Pion, G. M. , Brandt, C., & Molefy, A. (2011).*The Impact of the Measures of Academic Progress (MAP) Program on Student Reading Achievement: Final Report.* Unpublished report prepared for the Institute of Education Sciences under contract ED-06-CO-0019.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. Bootzin & P. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation*. Washington, DC: American Psychological Association.

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

De Boeck, P., & Leuven, K.U. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: a generalized linear and nonlinear approach.* New York: Springer.

Decker, G. (2003). Using data to drive student achievement in the classroom and on high–stakes tests. *THE Journal*. Retrieved August 3, 2012 from: http://thejournal.com/articles/2003/01/01/using-data-to–drive-student-achievement-in-the-classroom-and-on-highstakes-tests.aspx?sc_lang=en.

DeMars, 2001. Group Differences Based on IRT Scores: Does the Model Matter? *Educational and Psychological Measurement*, *61*, 60–70.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, *2*, 217–233.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (RR-83-9). Princeton NJ: Educational Testing Service.

Dorans, N. J., Schmitt, A. P. and Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*, 309–319.

Durlak, J.A., & DuPre, E.P. (2008). Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation. *American Journal of Community Psychology, 41*, 327–350.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W.B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*, 237–256.

Embretson, S. E. (1994). Comparing changes between groups: Some perplexities arising from psychometrics. In Laveault, D., Zumbo, B. D., Gessaroli, M. E., & Boss, M. W. (Eds.), *Modern theories of measurement: Problems and issues* (pp. 213–248). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*, 201–212.

Embretson, S. E., & Hershberger, S. (Eds.) (1999). *The new rules of measurement: What every psychologist and educator should know.* Mahwah, NJ: Erlbaum.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *33*, 291–314.

Field, A. & Miles, J. (2010). Discovering Statistics Using SAS. Thousand Oaks, CA: Sage Publications.

Fischer, G.H., & Molenaar, I.W. (1995). *Rasch models: Foundations, recent developments, and applications*. (1st ed.). Berlin, Germany: Springer-Verlag.

Fixsen, D.L., Naoom, S.F., Blasé, K.A., Friedman, R.M., Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature.* FMHI Publication no. 231. Tampa: Louis de la Parte Florida Mental Health Institute, National Implementation Research Network, University of South Florida.

Fraley, C.R., Waller, N.G. & Brennan, K.A. (2000). An Item Response Theory Analysis of Self-Report Measures of Adult Attachment. *Journal of Personality and Social Psychology*, *78*, 350–365.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239–256.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press.

Geisser, S., & Greenhouse, S. (1958). Extension of Box's results on the use of the F distribution in multivariate analysis. Annals of Mathematical Statistics, 29, 885–891.

Hall, T. (2002). Differentiated Instruction [Monograph]. Washington, DC, National Center on

Accessing the General Curriculum. Retrieved August 8, 2012 from: http://www.cast.org/system/galleries/download/ncac/DifInstruc.pdf.

Haggerty, K. P., Catalano, R. F., Harachi, T. W., & Abbott, R. D. (1998). Preventing adolescent problem behaviors: A comprehensive intervention description. *Criminologie, 31,* 25–47.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*, 38–47.

Harlow, L.L. (2005). *The essence of multivariate thinking: Basic themes and methods*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika, 74*, 191–210.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–960.

Hulleman, C.S., & Cordray, D.S. Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Intervention Effectiveness, 2*, 88–110.

IBM Corporation. (2011). *IBM SPSS Statistics.* Armonk, NY: IBM Corporation.

Jago, R., Baranowski, T., Baranowski, J.C., Thompson, D., Cullen, K.W., Watson, K., & Liu, Y. (2006). Fit for Life Boy Scout badge: Outcome evaluation of a troop and Internet intervention. *Preventative Medicine*, *42*, 181–187.

Jodoin, M. G., & Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329–349.

Jones, L.V. (1971). The nature of measurement. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education.

Kamata, A., & Bauer, D.J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15,* 136–153.

Kenny, D. A., & Judd, C. A. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*, 422–431.

Kirk, R.E. (1968). *Experimental Design: Procedures for the Behavioral Sciences*, Belmont, CA: Brooks Cole.

Knowlton, L.W., & Phillips, C.C. (2009). *The Logic Model Guidebook: Better Strategies for Great Results*. Washington, D.C.: Sage.

Levene, H. (1960). Robust Tests for Equality of Variances, in *Contributions to Probability and Statistics*, (1st ed.). Olkin, Palo Alto, CA: Stanford Univ. Press.

Linacre, J. K. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*, 328.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person parameters based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*, 921–943.

Mantel, N., & Haenszel, W., (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data*. Mahwah, NJ: Lawrence Erlbaum.

Mauchly, J.W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*, 204–209.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement,17,* 297–334.

Ndalichako, J. L., & Rogers,W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement*, *57*, 580–589.

Nelson, M.C., Cordray, D.S., Hulleman, C.S., Darrow, C.L., & Sommer, E.C. (in press). A Procedure for Assessing Intervention Fidelity in Experiments Testing Educational and Behavioral Interventions. *Journal of Behavioral Health Services & Research.*

Northwest Evaluation Association. (2003). *Technical manual for the NWEA Measures of Academic Progress and Achievement Level Tests.* Lake Oswego, OR: Author.

O'Donnell, C. L. (2008). Defining, Conceptualizing, and Measuring Fidelity of Implementation and Its Relationship to Outcomes in K–12 Curriculum Intervention Research. *Review of Educational Research*, *78*, 33–84.

Osteen, P. (2010). An Introduction to Using Multidimensional Item Response Theory to Assess Latent Factor Structures. *Journal of the Society for Social Work and Research*, *1*, 66–82.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.* Vienna, Austria. ISBN 3-900051-07-0, Retrieved on July 15, 2012 from http: //www.R–project.org/.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, *27*, 85–95.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207–320.

Reckase, M.D. (1997). The Past and Future of Multidimensional Item Response Theory. *Applied Psychological Measurement*, *21*, 25–36.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93–103.

Rijman, F. (2010). Formal Relations and an Empirical Comparison among the Bi–Factor, the Testlet, and a Second–Order Multidimensional IRT Model. *Journal of Educational Measurement*, *47*, 361–372.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 416–464.

Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test for normality. *Biometrika, 52,* 591–611.

University of Michigan School of Education. (2001). *Study of Instructional Improvement.* Retrieved on July 18, 2012 from http://www.sii.soe.umich.edu.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale NJ: Erlbaum.

Tomlinson, C. A. (2001) How *to Differentiate Instruction in Mixed Ability Classrooms.* Alexandria, Virginia: Association for Supervision and Curriculum Development.

Tomlinson, C. A. and McTighe, J. (2006) *Integrating differentiated instruction + understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

W. K. Kellogg Foundation. (2004). *Logic model developmentguide*. Retrieved April 12, 2011 from http://www.wkkf.org/~/media/36693510092544928C454B5778180D75/ LogicModel.pdf

Wang, W.C., Chen, P.H., & Cheng, Y.Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*, 116–136.

Wang, W., & Chyi–In, W. (2004), Gain score in item response theory as an effect size measure. *Educational and psychological measurement*, *64*, 758–780.

Watson, K., Baranowski, T., Thompson, D., Jago, R., Baranowski, J., & Klesges, L.M. (2006). Item response modeling: an evaluation of the children's fruit and vegetable self-efficacy questionnaire. *Health Education Research, 21*, 47–57.

Williams, Swanlund, Miller, Konstantopoulos, & van der Ploeg. (2012). Building measures of instructional differentiation from teacher checklists. Paper presented at the SREE 2012 Spring Conference. Evanston, IL: Society for Research on Educational Effectiveness.

Willse, J.T. & Goodman, J.T. (2008). Comparison of Multiple-Indicators, Multiple-Causes and Item Response Theory-Based Analyses of Subgroup Differences. *Educational and Psychological Measurement*, *68*, 587–602.

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research-Online*, *8*, 1–22.

Wu, Margaret L. (2007) *ACER ConQuest version 2.0: generalised item response modelling software.* ACER Press, Camberwell, Vic.

Wu, M. & Adams, R. (2006). *Journal of the Society for Social Work and Research*, *1*, 66–82.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, *47*, 339–360.

Yen, W.M. (1986). The choice of scale for educational measurement: an IRT perspective. *Journal of Educational Measurement*, *23*, 299–325.

Zellner, A. Estimation of Regression Relationships Containing Unobservable Variables. *Intemational Economic Review, 11*, 441–454.

Zickar, M.J., & Broadfoot, A.A. (2009). The partial revival of a dead horse? comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*. (pp. 37–59). New York, NY, US: Routledge/Taylor & Francis Group.

Zumbo, B. D., & Thomas, D. R. (1997) *A measure of effect size for a model–based approach for studying DIF*. Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.

Appendix A

Differentiated instruction subscales from the survey of teachers on instructional practices (Cordray et al., 2010).

Table A1: Subscale: Use of multiple instructional groups (IM)

#	Item	Description
1	Q10	Did you group students by ability or achievement for reading/language arts instruction?
2	Q10a	When teaching your reading class, how often did you group students by ability or achievement?
3	Q11	Once you assigned students to ability or achievement groups, approximately how often did you change the composition of these groups?
4	Q17	When you further grouped these high-achieving students for reading/language arts instruction according to their ability, approximately how large were these individual groups?
5	Q27	When you further grouped these low-achieving students for reading/language arts instruction according to their ability, approximately how large were these individual groups?

Table A2: Subscale: Diverse instructional strategies (DS)

"**Q19/Q29. Looking back over the school year and thinking about how you taught reading/language arts to these high/low-achieving students, how often did you…**"

#	Item	Description
6	a	Use basic skills worksheets
7	b	Use enrichment worksheets
8	c	Assign reading of more advanced level work
9	d	Assign reports
10	e	Assign projects or other work requiring extended time for students to complete
11	f	Make time available for students to pursue self-selected interests
12	g	Use pretests to determine if students had mastered the material covered in a particular unit or content area
13	h	Repeat instruction on the coverage of more difficult concepts for some students
14	i	Encourage students to move around the classroom to work in various locations
15	j	Allow students to leave the classroom to work in another location, such as the school library or media center
16	k	Use learning centers to reinforce basic skills
17	l	Use enrichment centers
18	m	Teach thinking skills such as critical thinking or creative problem-solving
19	n	Use contracts or management plans to help students organize their independent study projects
20	o	Establish interest groups which enable students to pursue individual or small group interests
21	p	Consider students' opinions in allocating time for various subjects within your classroom

Table A2, continued

"**Q19/Q29. Looking back over the school year and thinking about how you taught reading/language arts to these high/low-achieving students, how often did you…**"

#	Item	Description
22	q	Provide opportunities for students to use programmed or self-instructional materials at their own pace
23	r	Use computers
24	s	Encourage student participation in discussions

Table A3: Subscale: Diversity of instructional topics (IT)

"**Q18/Q28. How often were the following topics a primary focus of instruction for these high/low-achieving students?**"

#	Item	Description
25	a	Word analysis (e.g., decoding, word families, context cues, and sight words)
26	b	Reading fluency (e.g., repeated reading and guided oral reading)
27	c	Listening comprehension
28	d	Reading comprehension
29	e	Grammar
30	f	Spelling
31	g	Written composition (e.g., writing sentences, paragraphs, and stories)

Appendix B

R code used for IRT analyses

R Code for DIF detection methods:

```
##Loading R libraries
> library(Matrix)
> library(lattice)
> library(lme4)
> library(ltm)
> library(difR)

##Uploading data
> Cov <- read.table("DimensionWide.csv",sep=",",header=T,fill=T)

##Mantel-Haenszel method
> MH <- difMH(Cov, group="Covariate.var", focal.name=1)
> MH

##Standardized method
> SM <- difStd(Cov, group="Covariate.var", focal.name=1)
> SM

##Logistic regression for uniform and non-uniform DIF
> LR <- difLogistic(Cov, group="Covariate.var", focal.name=1)
> LR

##Logistic regression for uniform DIF only
```

```
> LR <- difLogistic(Cov, group="Covariate.var", focal.name=1,
type="udif")

> LR


##Logistic regression for non-uniform DIF only

> LR <- difLogistic(Cov, group="Covariate.var", focal.name=1,
type="nudif")

> LR


##Likelihood ratio test

> LRT <- difLRT(Cov, group="Covariate.var", focal.name=1,
purify=TRUE, nrIter=10)

> LRT
```

R Code for descriptive IRT models:

```
##Uploading data


> DI.long <- read.table("DI.txt",header=T,fill=T)

> DI.long$Dimension=factor(DI.long[,31])

> DI.long$Grade.var=factor(DI.long[,3])

> DI.long$Role.var=factor(DI.long[,2])


##Unidimensional model


> UDM <- lmer(Y ~ -1 + Item.01 + Item.02 + Item.03 + Item.04 +
Item.05 + Item.06 + Item.07 + Item.08 + Item.09 + Item.10 +
Item.11 + Item.12 + Item.13 + Item.14 + Item.15 + Item.16 +
Item.17 + Item.18 + Item.19 + Item.20 + Item.21 + Item.22 +
Item.23 + (1|Person), DI.long, binomial("logit"))
```

```
>UDM


##Consecutive models


> DI.long.im <- data.frame(subset(DI.long, IM.var==1))

> DI.long.ds <- data.frame(subset(DI.long, DS.var==1))

> DI.long.it <- data.frame(subset(DI.long, IT.var==1))

> imirt <- lmer(Y ~ -1 + Item.01 + Item.02 + Item.03 + Item.04 +
Item.05 + (1|Person), DI.long.im, binomial("logit"))

> imirt

> itirt <- lmer(Y ~ -1 + Item.06 + Item.07 + Item.08 + Item.09 +
Item.10 + Item.11 + Item.12 + (1|Person), DI.long.it,
binomial("logit"))

> itirt

> dsirt <- lmer(Y ~ -1 + Item.13 + Item.14 + Item.15 + Item.16 +
Item.17 + Item.18 + Item.19 + Item.20 + Item.21 + Item.22 +
Item.23 + (1|Person), DI.long.ds, binomial("logit"))

> dsirt


## Descriptive multidimensional model


> MDM1 <- lmer(Y ~ -1 + Item.01 + Item.02 + Item.03 + Item.04 +
Item.05 + Item.06 + Item.07 + Item.08 + Item.09 + Item.10 +
Item.11 + Item.12 + Item.13 + Item.14 + Item.15 + Item.16 +
Item.17 + Item.18 + Item.19 + Item.20 + Item.21 + Item.22 +
Item.23 + (IM.var + DS.var + IT.var - 1|Person), DI.long,
binomial("logit"))

> MDM1
```

```
##R code for explanatory multidimensional model:
> MDM2 <- lmer(Y ~ 1 + Dimension*Grade.var + Dimension*Role.var
+ Dimension*Grade.var*Role.var + (IM.var + DS.var + IT.var -
1|Person) + (1|Item), DI.long, binomial("logit"))
```

Appendix C

Differential item functioning results for all items

Key for all tables:

ETS delta effect size codes: A = 0.0 to 1.0, B = 1.0 to 1.5, C = 1.5+

Dorans, Schmitt & Bleistein effect size codes: A = 0.00 to 0.05, B = 0.05 to 1.00, C = 1.00+

Zumbo & Thomas effect size codes: A = 0.00 to 0.13, B = 0.13 to 0.26, C = 0.26 to 1.00

Jodoign & Gierl effect size codes: A = 0.00 to 0.035, B = 0.035 to 0.07, C = 0.07 to 1.00

p-values (alpha = 0.05): * = 0.01 to 0.05, *** = 0.000 to 0.001

Table C1: Instructional Modality x Experimental Condition DIF effect sizes:

	MH ES	**Standardized DSB ES**	**LR Both ZT/JG**	**LR Uniform ZT/JG**	**LR Non-Uniform ZT/JG**	**Like. Ratio *p*-value**
1	C	A/A	A/C	A/A	A/C	.66
2	A	A/A	A/A	A/A	A/A	.53
3	A	A/A	A/A	A/A	A/A	.65
4	A	A/A	A/A	A/A	A/A	.64
5	A	A/A	A/C	A/A	A/C	.74

Table C2: Instructional Modality x Grade Level DIF effect sizes:

	MH ES	Standardized DSB	LR Both ZT/JG	LR Uniform ZT/JG	LR Non-Uniform ZT/JG	Like. Ratio *p*-value
1	B	A/B	A/B	A/A	A/B	.80
2	A	A/A	A/A	A/A	A/A	.95
3	B	B/A	A/A	A/A	A/A	.49
4	C	B/B	A/A	A/A	A/A	.50
5	B	A/A	A/A	A/A	A/A	.72

Table C3: Diverse Instructional Strategies x Experimental Condition DIF effect sizes:

	MH ES	Standardized DSB ES	LR Both ZT/JG	LR Uniform ZT/JG	LR Non-Uniform ZT/JG	Like. Ratio *p*-value
6	C	C/C	B/C	B/C	A/A	< .01***
7	C	C/C	B/C	A/C	A/A	.01*
8	A	B/A	A/A	A/A	A/A	.54
9	C	A/C	A/B	A/B	A/A	.05*
10	A	A/A	A/B	A/A	A/B	.71
11	C	C/C	A/B	A/B	A/A	.19

Table C3, continued:

	MH ES	Standardized DSB ES	LR Both ZT/JG	LR Uniform ZT/JG	LR Non-Uniform ZT/JG	Like. Ratio *p*-value
12	A	A/A	A/A	A/A	A/A	.85
13	A	B/A	A/A	A/A	A/A	.99
14	B	A/A	A/A	A/A	A/A	.58
15	C	B/A	A/A	A/A	A/A	.32
16	A	A/A	A/A	A/A	A/A	.36
17	C	A/B	A/A	A/A	A/A	.56
18	B	C/B	A/A	A/A	A/A	.64
19	A	A/A	A/A	A/A	A/A	.55
20	A	A/A	A/A	A/A	A/A	.82
21	A	B/B	A/A	A/A	A/A	.55
22	A	A/A	A/A	A/A	A/A	.62
23	C	C/B	A/B	A/B	A/A	.04*
24	C	A/B	A/A	A/A	A/A	.26

Table C4: Diverse Instructional Strategies x Grade Level DIF effect sizes:

	MH ES	Standardized DSB ES	LR Both ZT/JG	LR Uniform ZT/JG	LR Non-Uniform ZT/JG	Like. Ratio *p*-value
6	A	C/B	A/A	A/A	A/A	.33
7	A	A/A	A/B	A/A	A/A	.51
8	C	B/A	A/C	A/A	A/B	.17
9	C	A/C	A/C	A/C	A/B	.67
10	C	C/C	A/B	A/B	A/A	.29
11	A	B/A	B/C	A/A	B/C	.45
12	C	A/C	A/B	A/B	A/A	.33
13	A	A/A	A/B	A/A	A/B	.97
14	A	C/B	A/A	A/A	A/A	.93
15	A	B/A	A/A	A/A	A/A	.75
16	A	A/A	A/A	A/A	A/A	.66
17	C	C/C	B/C	A/C	A/B	< .01***
18	C	B/A	A/B	A/A	A/A	.15
19	A	A/A	A/A	A/A	A/A	.97
20	C	C/C	A/A	A/A	A/A	.67
21	A	A/A	A/A	A/A	A/A	.76
22	B	A/A	A/A	A/A	A/A	.08
23	A	A/A	A/A	A/A	A/A	.90
24	C	B/C	A/B	A/A	A/B	.26

Table C5: Instructional Topics x Experimental Condition DIF effect sizes:

	MH ES	Standardized DSB ES	LR Both ZT/JG	LR Uniform ZT/JG	LR Non-Uniform ZT/JG	Like. Ratio *p*-value
25	A	A/A	A/A	A/A	A/A	.49
26	A	A/A	A/A	A/A	A/A	.59
27	C	B/A	A/A	A/A	A/A	.07
28	C	A/A	A/A	A/A	A/A	.26
29	A	A/A	A/A	A/A	A/A	.28
30	B	B/A	A/A	A/A	A/A	.19
31	A	A/A	A/A	A/A	A/A	.92

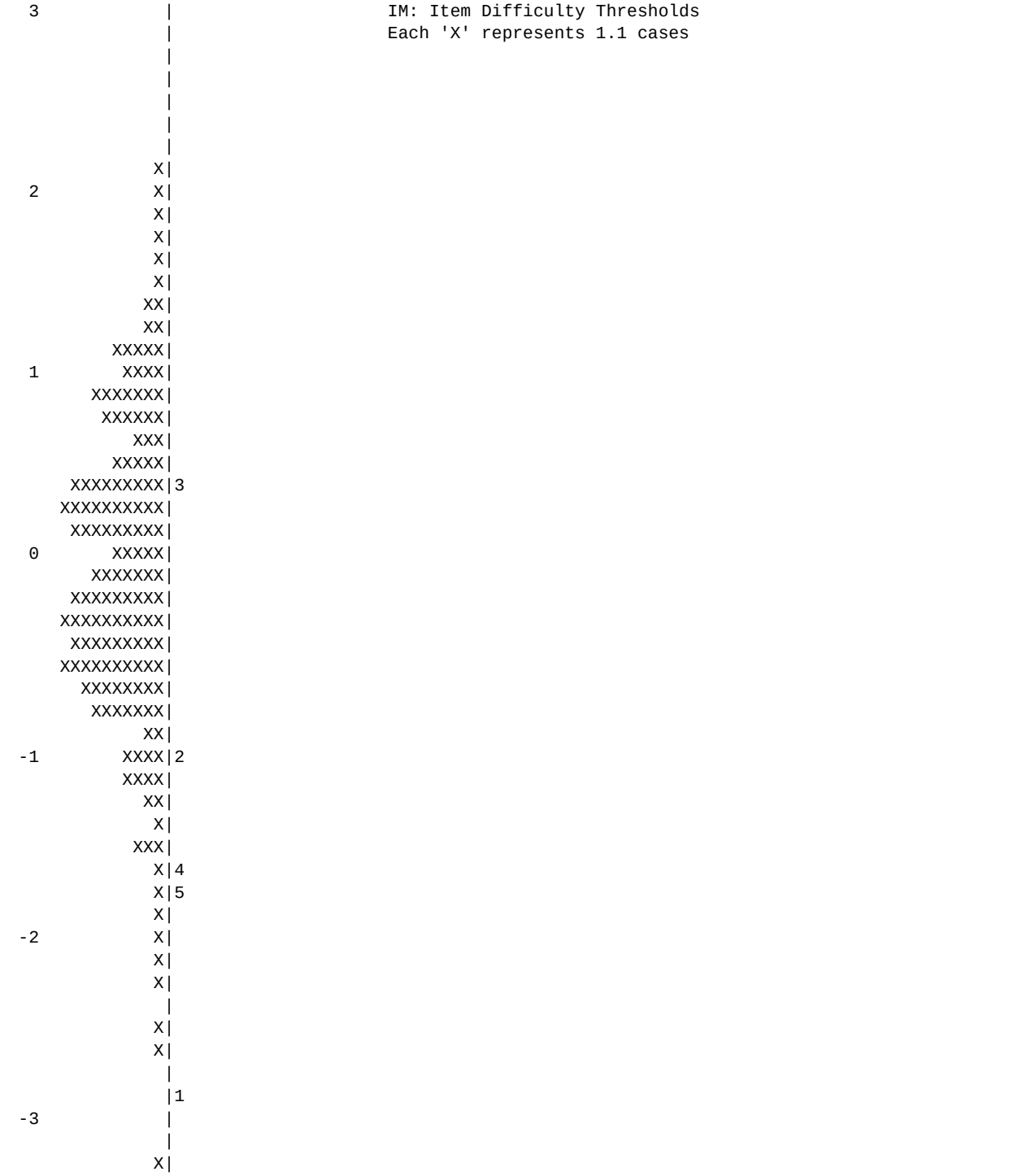
Table C6: Instructional Topics x Grade Level DIF effect sizes:

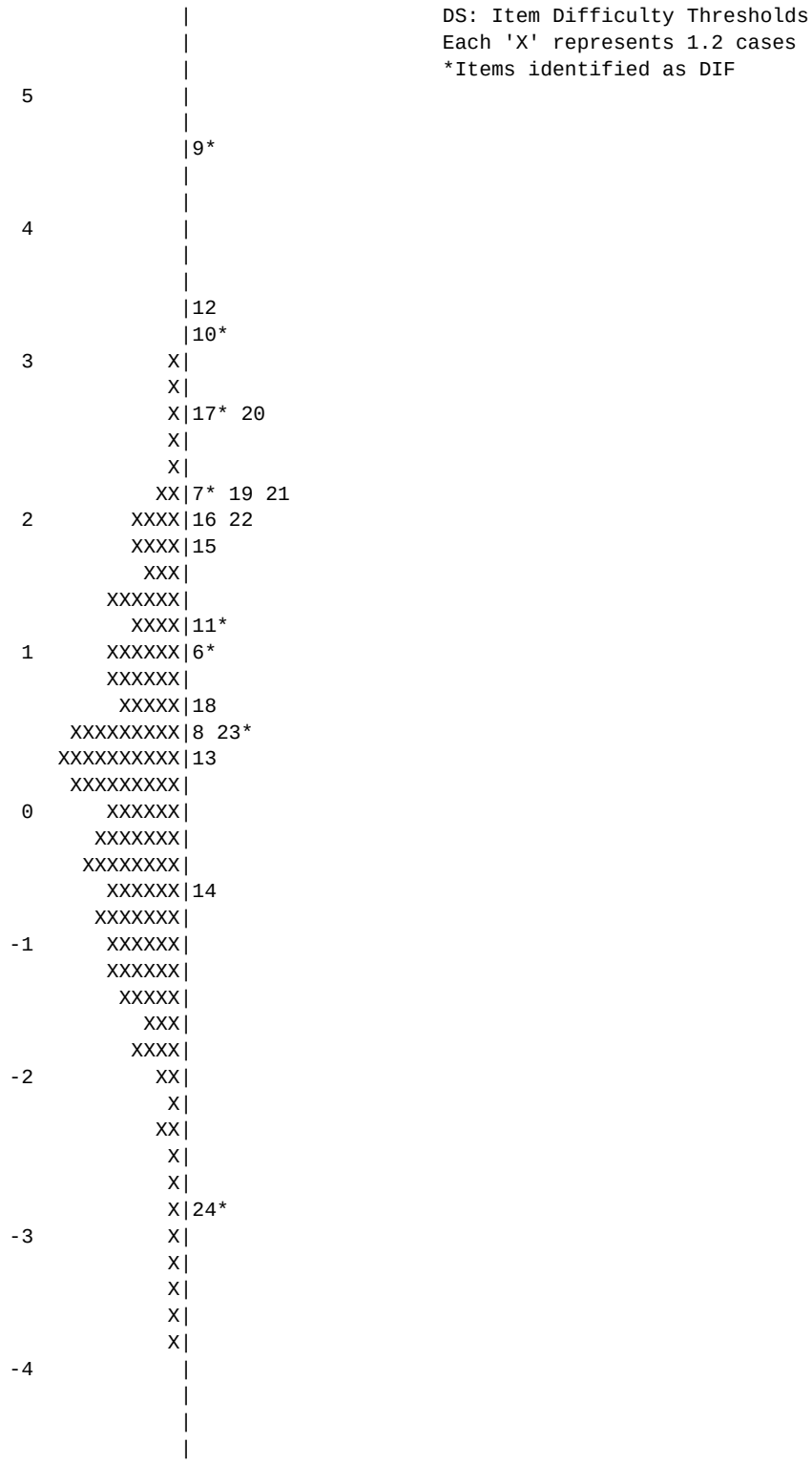
	MH ES	Standardized DSB ES	LR Both ZT/JG	LR Uniform ZT/JG	LR Non-Uniform ZT/JG	Like. Ratio *p*-value
25	C	C/B	A/B	A/A	A/A	.08
26	A	A/A	A/A	A/A	A/A	.56
27	A	A/A	A/A	A/A	A/A	.87
28	C	A/B	A/A	A/A	A/A	.11
29	A	A/A	A/A	A/A	A/A	.87
30	C	C/A	A/B	A/A	A/B	.06
31	C	B/A	A/A	A/A	A/A	.10

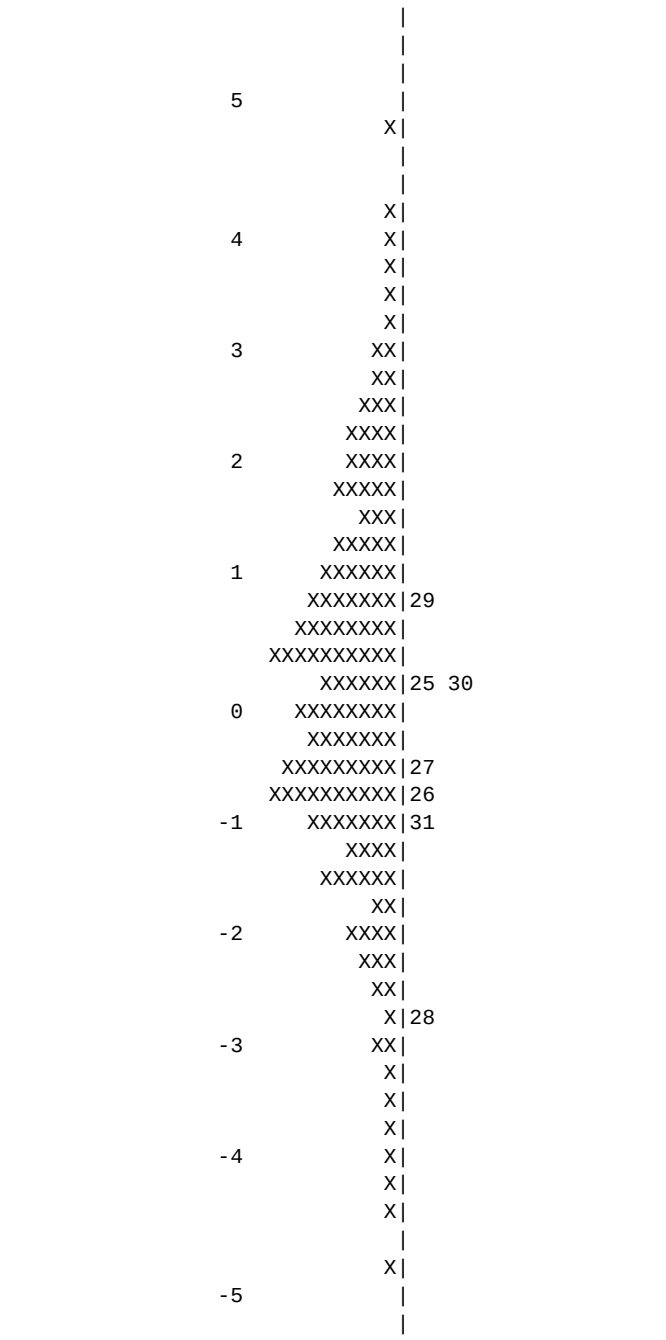
Appendix D

Subscale Wright maps

Scales are in logits, higher values representing higher trait values (left) or item difficulty (right).




DS: Item Difficulty Thresholds
Each 'X' represents 1.2 cases
*Items identified as DIF
5
4
3
2
1
0
-1
-2
-3
-4
9*
12
10*
17* 20
7* 19 21
16 22
15
11*
6*
18
8 23*
13
14
24*


IT: Item Difficulty Thresholds
Each 'X' represents 1.2 cases
5
4
3
2
1
0
-1
-2
-3
-4
-5
29
25 30
27
26
31
28