Challenges & Successes of Building Measures Relevant to Research on Teacher Education in International Contexts

Michael C. Rodriguez, University of Minnesota

April, 2012

Paper presented in the symposium

"Measuring the impact of teacher education on learning to teach mathematics:
The teacher education study in mathematics"
(Organizer, Maria Teresa Tatto)

at the annual meeting of the
American Educational Research Association,
Vancouver, British Columbia.

Introduction

The Teacher Education Study in Mathematics (TEDS-M) is an ambitious international comparative study of future mathematics teachers in their final year of teacher preparation, their educators, program characteristics, and teacher preparation policy environments. The study, directed by Michigan State University in collaboration with the Australian Council for Educational Research and national research centers in 17 countries, was supported through the International Association for the Evaluation of Educational Achievement (IEA) with funding from the National Science Foundation. Comprehensive information about the TEDS-M study can be found in the conceptual framework (Tatto, Schwille, Senk, Ingvarson, Peck, & Rowley, 2008) and the assessment frameworks (Tatto, Senk, Bankov, Rodriguez, & Peck, 2011). Initial findings from the study have been published recently (Tatto, Schwille, Senk, Ingvarson, Rowley, Peck, Bankov, Rodriguez, Reckase, 2012; Tatto & Senk, 2011).

The study was comprised of several components, including a curriculum and syllabi analysis; surveys, document reviews, and interviews regarding teacher preparation policies; and a trio of surveys including a large-scale survey of a probability-based cluster sample of future teachers of mathematics, a survey of their educators, and a survey of teacher-preparation program characteristics. The design of the TEDS-M instruments took several phases, including an early item tryout, a formal pilot, and operational instrument design. A formal process was developed to ensure coherence and consistency in item format and presentation for all items in all the questionnaires and on the mathematics knowledge test items.

The presentation here is primarily methodological concerning the design and scaling of the future teacher survey, including the application of modern measurement theories to solve practical problems with large-scale international survey research. This includes latent-trait methods such as confirmatory factor analysis using Mplus (Muthén & Muthén, 2007) and the Rasch scaling model using Winsteps (Linacre, 2009). In the process of applying latent-trait methods to the survey items, challenges arose that complicated the analyses, introducing complexities in the testing of assumptions and the provision of strong validity evidence. These challenges were partly a function of the complex teacher preparation program ecologies (heterogeneous models of teacher preparation across countries) and the inherent complexity introduced through latent-trait models and resulting challenges in interpreting measures in useful and meaningful ways. Resolutions to these challenges are offered.

Instrumentation

Two broad areas of characteristics of future teachers and teacher preparation programs were identified as integral to the conceptual framework for TEDS-M, including opportunity to learn (OTL) and beliefs about teaching and learning mathematics (beliefs). OTL was conceived of in four areas, including (a) mathematics content, (b) mathematics education pedagogy, (c) general education pedagogy, and (d) school-based experiences. Beliefs were conceptualized in three areas, including (e) beliefs about the nature of mathematics, (f) beliefs about learning mathematics, and (g) beliefs about mathematics achievement. Three more general concepts were relevant to (h) teacher preparation programs as a whole included program coherence, instructional quality, and preparedness for teaching mathematics. The specific titles of the OTL and beliefs measures are listed in Table 1.

Table 1
TEDS-M Measures of Opportunities to Learn and Beliefs about Teaching and Learning

Category	Index/measure	Future teachers	Educators
Mathematics content			
Tertiary level mathematics	Geometry	MFB1GEOM	-
	Discrete Structures & Logic	MFB1DISC	-
	Continuity & Functions	MFB1CONT	-
	Probability & Statistics	MFB1PRST	-
School Level Math	Numbers Measurement Geometry	MFB2SLMN	-
	Functions Probability Calculus	MFB2SLMF	-
Mathematics education pedagogy	Foundations of math ed pedagogy	MFB4FOUN	-
	Principles of instruction and curriculum	MFB4INST	-
	Engage in class participation	MFB5PART	MEI1PART
	Engage in class readings	MFB5READ	MEI1READ
	Engage in solving problems in class	MFB5SOLV	MEI5SOLV
	Instructional practice and techniques	MFB6IPRA	MEG2IPRA
	Instructional planning techniques	MFB6IPLA	MEI3IPLA
	Uses of assessment	MFB6AUSE	MEI3AUSE
	Assessment practices and techniques	MFB6APRA	MEG2APRA
Education pedagogy	Social sciences (history, philosophy, sociology)	MFB7EPSS	-
	Application of general education pedagogy	MFB7EPAP	-
	Teaching for diversity	MFB8DVRS	MEH2DVRS
	Teaching for reflection on practice	MFB8REFL	MEH2REFL
	Teaching for improving practice	MFB9IMPR	MEH1IMPR
School-based experience	Connecting classroom learning to practice	MFB13CLP	MEI2CLP
	Supervising teacher reinforcement of university goals for practicum	MFB14STR	-
	Supervising teacher feedback quality	MFB14STF	-
The nature of mathematics	Mathematics is a set of rules and procedures	MFD1RULE	MEK1RULE
	Mathematics is a process of inquiry	MFD1PROC	MEK1PROC
Learning mathematics	Mathematics is learned through teacher direction	MFD2TEAC	MEK2TEAC
	Mathematics is learned through active learning	MFD2ACTV	MEK2ACTV
Mathematics achievement	Mathematics is a fixed ability	MFD3FIXD	MEK3FIXD
The program as a whole	Program coherence	MFB15COH	MEJ1COH
	Instructional quality	MFD5QUAL	-
	Preparedness for teaching mathematics	MFD4PREP	MEL1PREP

Note. The variable labels used in the TEDS-M database are listed under the Future teachers and Educators columns. Not all OTL measures are used with educators as they include student-specific experiences not relevant to the experiences of Educators.

Items were solicited from the participating international research teams and many were developed to reflect the principles of mathematics knowledge for teaching. Some items were modeled after other international educational surveys, such as TIMSS, and from an earlier small-scale version of TEDS-M. Questions for opportunities to learn and beliefs were also obtained from related research instruments developed in the USA and Australia through the Australian Council for Educational Research, with input from other countries. All items were reviewed by all participating countries and the TEDS-M management team.

Initial development occurred prior to TEDS-M through a series of preliminary research efforts (knows as Pre-TEDS) and others developing items for unrelated projects. Two such scales, *Connecting Theories of Teaching & Learning* and *Connecting Practice and Reflection*, were developed by the Australian Council for Educational Research. Many items were offered for consideration to the TEDS-M Management team. From the large body of items, many more than could be used in the operational survey were piloted in forms administered to participating countries, to gather evidence of item quality. With the pilot data obtained in 2007, initial exploratory factor analyses were conducted to identify homogenous item sets covering the core measures of OTL and beliefs that also functioned similarly for primary and secondary education future teachers. Briefly, correlations between items and total scores (item discrimination) were examined across each measure by country to select those items with the strongest indicators of commonalities across countries. The pilot was based on relatively small samples within some countries, so a full factor analysis by country model was not possible.

Generally speaking, the items within measures worked as expected and measures identified through factor analyses were consistent with prior research. As will be described below, the factor structures identified through the pilot (which were consistent with prior findings) were remarkably consistent with the factor structures resulting from the operational data – confirmed through confirmatory factor analyses.

General Scaling Methods

In all stages of the TEDS-M item development, item analyses were conducted, including exploratory factor analysis, and correlations among and between items of similar and different constructs. These data were used in conjunction with careful review of content to inform decision making about the final set of items included in operational measures. The items functioned exceptionally well, as expected from prior experience with pre-existing items. Many items underwent revision based on item pilot reviews. A standard set of item-writing guidelines was adopted to ensure consistency and coherence in all measurement aspects. For the operational survey data, confirmatory factor analyses were conducted prior to scaling (employing Mplus), including analyses of factor structures across participating countries. In a small number of cases, this led to the elimination of a handful of items from the scaling procedure for some indicators of OTL and beliefs, particularly regarding those items that functioned significantly different in some countries. Once the final scales were defined, the Rasch model was used to create score scales for study participants on the measures of OTL and beliefs (a similar process was used to scale the knowledge measures, which is not described here). The Rasch analyses also provided additional scale quality information described below.

Many steps along the way provided useful and interesting information regarding the use of the many various measures as tools for understanding the context and outcomes of teacher education internationally. These processes were challenging at times, and in some cases planned

procedures were not possible given disjuncture between data characteristics, assumptions, and model theory. For example, the use of a confirmatory factor analysis approach to asses measurement invariance across countries (Does the scale measure the same thing in every country?) was not possible because of idiosyncratic response patterns in some countries. The challenges and successes in building relevant measures are presented here. There are primarily three major areas of challenges in developing measures of OTL and beliefs in an international context regarding preparation to teach mathematics. These are the three major areas presented in this paper, with specific examples and solutions (compromises), including:

1. Confirmatory Factor Analysis and the challenge of assessing measurement invariance across countries with complex and heterogeneous teacher preparation ecologies;
2. Rasch scaling and the challenge of the arbitrary scale metric;
3. The use of indicators measured at the individual level and the extent to which they are indicators of individual experiences versus program characteristics.

A Validity Framework

A core element of any study is measurement quality, which is required to support inferences from any measure. Validity is a key indicator of measurement quality. Current definitions of validity vary across fields; however in educational testing, most employ the framework described in the *Standards for Educational and Psychological Testing* (hereafter referred to as *Testing Standards*; AERA, APA, NCME, 1999). "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, NCME, 1999, p. 9). The *Testing Standards* describes validation as the process of gathering evidence to achieve these goals, including evidence related to the construct, test content, response processes, internal structure, and relations to other variables.

In all cases, validation is an ongoing process and the most important sources of validity evidence are those that are most closely related to the immediate inferences and proposed claims we make regarding measurement results. What evidence do we need to support the intended meaning of TEDS-M results? The primary inferences intended from the measures of OTL and beliefs focus on the content of those measures, particularly regarding OTL. The very idea of OTL presents a challenge to measurement (more on this in a moment). Another core inference regards the appropriateness, usefulness, and meaningfulness of each measure across international contexts, such that comparative analyses and inferences can be defended. In this sense, the question of measurement invariance is critical: Are the measures functioning similarly across countries? The evidence we are able to assemble in this respect includes content-related validity evidence, evidence of the internal structure of the measures and its consistency across contexts, and relations to other variables (as this is the key function of the conceptual framework, as OTL and beliefs are viewed to be critical moderators (and in some respects outcomes) of teacher preparation programs regarding teacher effectiveness.

Content-related validity evidence is found in the TEDS-M conceptual framework, the assessment frameworks, and the technical manual. This evidence is found directly in the items themselves. The specific items used to create each measure are also available in the database and others are able to evaluate the functioning of these items as indicators of each measure. The provision of this evidence is the responsibility of the measure developer.

Evidence of relations to other variables is currently under investigation, as this is the primary function of the TEDS-M study. Many researchers will use the TEDS-M database to

evaluate the associations among OTL and beliefs measures, knowledge measures, and many other background variables, characteristics of programs, and feedback from educators, to understand variation in country teacher preparation programs. The extent to which measures of OTL and beliefs function as intended in the conceptual framework provides additional validity-related evidence.

Additional important sources of validity evidence that are the responsibility of the measure developer is construct-related evidence and evidence regarding its internal structure. In particular, because of the international comparative nature of the intended inferences, evidence regarding measurement invariance is important to provide (challenge 1). This also requires scales that are appropriate, meaningful, and useful (challenge 2). Finally, the extent to which measures that are obtained at the individual level are really indicators of individual characteristics or program characteristics determines the level at which inferences can be appropriately made (challenge 3).

Challenge 1: CFA & Measurement Invariance

Confirmatory factor analysis (CFA) is a statistical technique that allows us to assess the internal structure of a measure, based on the common factor model. Each indicator or item (survey question) is a linear function of one or more factors (components that are common among the items) and one unique factor (the piece that is unique to each individual item). CFA partitions item variance into (a) common variance that is accounted for by the latent factor and (b) unique variance specific to the item as well as random error variance (measurement error). In the development of a high quality measure, we hope that most of the variance among a set of items is common variance, as the items are intended to be strong indicators of the common latent factor. Because there are statistical tests of model-data fit and indicators of the degree to which each item contributes to the common latent factor (factor loadings), CFA not only assesses the strength of the internal structure of a measure but can provide construct-related validity evidence as a test of whether the construct as defined by the researcher can be measured in a consistent way with a given set of items (the items contribute to a consistent score in a scale consisting of items that are arguably represent the construct). This also speaks to the reliability of the resulting scores, where in the modern conceptualization of validity, score reliability is validity evidence.

Measurement Model Considerations

Score reliability is an index of the degree to which the score captures common variance versus unique variance (as an index of the ratio of measurement error; see Haertel, 2006 for a comprehensive description of score reliability). However, how we define measurement error (and thus the composite score) is not an easy task – it is not a simple sum of item unique variances; it depends on the nature of the measurement model and how items are allowed to function in that model. There are four common measurement models used to describe the function of items (see Graham, 2006, for a complete description of these measurement models). Briefly, the parallel indicator model requires that all observed items measure the same latent factor on the same scale with the same degree of precision and amount of error (have equal factor loadings and equal error variances) so that parallel indicators measure the latent construct equivalently. The tau-equivalent model is similar to the parallel (most restrictive) model, except that the item error variances are free to differ. Tau equivalence is the assumption underlying the

derivation and interpretation of coefficient alpha, the most commonly reported form of score reliability. Essentially tau-equivalent indicators have levels of precision that can differ by a constant across pairs of indicators, with different error variances – so the item means are affected, not the item variances (scale) or covariance with other items. Finally, congeneric indicators (the least restrictive model) are presumed to measure the same construct (as in the case of all the other models), but the size of the factor loadings (scale) and uniquenesses (measurement error) are free to vary, where the assumption of independent measurement error must hold. In a typical CFA, the latent factor scale is set by fixing the path of one indicator to 1 and allowing all other paths to be freely estimated, as well as the error terms and error variances; this is a congeneric model. Since this allows the item variances and covariances to vary, this directly impacts the estimation of reliability. The point here is that the way we define measurement error (given the appropriate measurement model) determines how score reliability should be estimated and interpreted.

Consider a practical example regarding essentially tau-equivalent models. Two items from the measure of program coherence are:

B15e. *Each of my courses was clearly designed to prepare me to meet a common set of explicit standard expectations for beginning teachers*.

B15f. *There were clear links between most of the courses in my teacher education program*.

The items are on the standard 4-point scale: disagree, slightly disagree, slightly agree, agree. Both of these items speak to the internal coherence of the teacher education program (measure the same latent factor on the same scale), but 15e is much more restrictive, requiring that coherence to be true of ***each*** (every) course, whereas 15f speaks of ***most*** courses. Assuming that the item variances are the same across all items in the measure, we can say they all measure coherence on the same scale; however precision in measuring coherence will differ. If this model holds, differences in the item means do not affect the variance components used to estimate reliability.

Consider an example of the congeneric model. In the example of items B15e and B15f above, if the variances of the items differed (which is perhaps the most typical case in real data), the more appropriate measurement model is one which allows item variances (and error variances) to vary across items. Different item variances result when distributions across the rating scale differs from item to item. In the case of the coherence items, the replicate weighted item statistics for the Secondary Future Teacher sample (weighted N = 21,157) shows us that item means vary from 3.0 to 3.2 and item variances range from 0.65 to 0.87. Considering another measure, Beliefs about the nature of mathematics (Questions D1a-D1h) on a 6-point scale from strongly disagree to strongly agree, the replicate weighted item statistics for Secondary Future Teacher sample shows us that item means vary 2.6 to 4.1 and item variances range from 1.7 to 2.3. These results, and those from the more formal tests of the CFA, suggest that the congeneric measurement model is the most appropriate from which to estimate reliability and understand the internal structure and functioning of the items and overall measures.

Complications in Modeling Measures

Throughout the survey and across countries, there are very different levels of missing responses to specific items. In some countries, the degree of missingness is substantial, as respondents tended to not complete the survey – which was admittedly very long. Part A

(background) required 31 responses, Part B (OTL) required 148 responses, Part C (knowledge) required 43 responses (which includes 7 constructed responses), and Part D (beliefs) required 43 responses. It was anticipated to take participants 90 minutes to complete the survey.

A decision was required regarding the amount of missingness that would be allowable to estimate total scores on the measures of OTL and Beliefs. Given the small number of items for some of the measures, the management team decided participants must have responded to at least half of the items to be included in the scoring and reporting of OTL and Beliefs measures.

Because of the selection rule for inclusion in scaling, the weights had to be adjusted or "normalized" to account for the shift in sample sizes due to the selection rule. Weights were proportionally adjusted so that the effective sample sizes within each country were weighted to approximate the population sizes appropriately. Similarly, sample sizes across countries varied greatly; for example at the primary education level, samples of future teachers ranged from less than 100 in one country to well over 2000 in another. To facilitate the CFA modeling and recognize the unequal sample sizes (population sizes) across countries, the weights were further adjusted to balance the impact of each country in the CFA modeling – each country was then weighted to have a population size of 300. This balanced the issues of case selection and influence of country size in the resulting CFA.

Measurement Invariance

Group comparisons of latent means are meaningful only if the factor loadings and indicator intercepts have been found to be invariant across groups, a condition referred to as measurement invariance. Group comparisons of factor variances and covariances (associations between factors) are meaningful when factor loadings are invariant (Brown, p. 269). In the tradition of analyses of measurement invariance, there is a series of tests that successively constrain parameters of the model, each testing a stricter level of invariance. This process begins with a simple test of CFA across all groups. This was described above. The next step is to test the CFA model separately for each group. This was also accomplished and the fit statistics for each CFA model by country are reported in the TEDS-M Technical Manual.

The following steps in the analysis of measurement invariance begins with a test of configural invariance (weak factorial invariance) where the same pattern of fixed and free factor loadings is specified for each group (Horn & McArdle, 1992). This first test is evaluating a hypothesis of invariant congeneric measurement properties across groups (Vandenberg & Lance, 2000). By achieving this level, it suggests that participants employ the same conceptual frame of reference to the set of items (essentially, the factor structure is the same across groups; for example, the items comprise a two-factor structure for all groups, rather than a three-factor structure for some groups).

Further tests of measurement invariance (as described by Brown, 2006) require constraints on the CFA model including the test of equality of factor loadings (where factor loadings are constrained to equality across groups, a form called metric invariance); the equality of indicator intercepts (where intercepts are constrained to equality across groups allowing comparison of mean differences across groups, a form called scalar invariance), and the equality of indicator residual variances (where residual variances are constrained to equality across groups, a form called strict factorial invariance, which some argue is optional since it is nearly impossible to achieve in practice).

Several complications arose in the first step of examining metric invariance, largely due to the complex nature of the teacher preparation ecologies and varied characteristics of teacher preparation programs. Byrne, Shavelson, & Muthén (1989) suggested that invariance evaluation may proceed in the context of partial measurement invariance. This is supported by providing substantive accounts for the sources of noninvariance. Here we can introduce some of the conditions that account for noninvariance. Using a chi-square difference test between nested models (models with successively restrictive constraints), the null-hypothesis is "the model fit does not get worse through the introduction of invariance constraints."

We note three features of the data that make measurement invariance testing difficult in this context. First, model fit is generally more difficult with more complex models, and in the TEDS-M multiple group CFA, we are comparing model parameters across 15 countries at the primary and secondary education levels. This alone is a difficult task to achieve.

A second complication is that factor structures will vary explicitly as a function of the response distributions to each item across countries. For example, for the school-based experience measure of Connecting Classroom Learning to Practice, none of the participants in the Philippines reported to never "Practice theories for teaching mathematics that you were learning in class." This resulted in the item consisting of three levels (three-point rating scale), which functionally makes this item and its factor have a different structure.

Another example is with the measures of school-based experience concerning supervising teachers. Again, in the Philippines, we find no participants disagreeing with (Question B14a) "I had a clear understanding of what my school-based supervising teacher expected of me…", (Question B14f) "The feedback I received from my supervising teacher helped me to improve my understanding of pupils", (Question B14G) "The feedback I received from my supervising teacher helped me to improve my teaching methods", and (Question B14i) "The feedback I received from my supervising teacher helped me improve my knowledge of mathematics content." Similarly in Thailand, no participants disagreed with question B14a.

The complexity of evaluating measurement invariance in many varied contexts that are also culturally different is daunting. In cases, programs focus on mathematics content, in others mathematics education pedagogy, in others general education pedagogy, and programs differ in their emphasis of school-based experiences. However, the structure of the measure and its consistent meaning across context should not be affected. But when participants in one country do not use all levels of a rating scale, empirically the item then has a different scale structure.

The complexity of measurement invariance introduces a few challenging considerations. If measurement invariance does not hold, it might suggest that the set of items does not represent the construct domain as conceptualized by each group. It is also possible that the construct is similar, but the cognitive frame of reference differs across groups; however, the current tests may not be sensitive enough to test such propositions (Vandenberg, 2002). Vandenberg (2002) also suggested that certain triggers can interfere with positive invariance results, including cultural issues – for example individuals from individualistic versus collectivistic cultures employ different frames of reference, resulting in very different response patterns to items sensitive to such cultural differences. In other cases, he argues that interventions or training might create shifts in frames of reference that result in failure to observe invariance.

There are conditions for pursuing meaningful group comparisons under partial invariance, where at least some of the parameters at each level of constraints are invariant (Brown, 2006; Vandenberg & Lance, 2000). Further research should be done to evaluate the extent to which partial invariance might be sought to support meaningful international

comparative hypotheses. Finally, for those items contributing to noninvariance, differences might provide meaningful information regarding relevant characteristics of culture and/or preparation programs leading to variation in the usefulness of common indicators of OTL or beliefs.

Challenge 2: Rasch Scaling and Scale Meaning

Rasch scaling was used to produce the reporting score scale for the OTL and beliefs indices. There is a long tradition of the use of Rasch scaling in education, health research, psychology, marketing, economics, social sciences, and even in the ranking of sports teams and players. Rasch scaling provides score scales for OTL and beliefs that have several scale (statistical) properties that make them stronger variables in General Linear Model (GLM) based analyses. When the assumptions of the model are met (unidimensionality and local independence of item responses), Rasch scales result in interval-level measurement (Harwell & Gatti, 2001), providing a scale with properties suited for correlational methods. The improved scale properties relative to the use of a simple summed score is probably the most significant benefit of using Rasch scaling. The Rasch analysis locates each item (actually, each point on the rating scale within the item) on the same scale as person trait levels, providing for a meaningful ordering of indicators relaying information about the rarity or severity of each indicator (a form of item difficulty). The Rasch scaling provides an efficient way to estimate trait values for individuals who have not responded to every item. However, it is important to note that OTL indices as conceptualized in the study framework are indicators of program characteristics, and as such, are to be used in their aggregate at the program level. Person-trait levels, as estimated by Rasch, are useful in this context as indicators of program characteristics. The extent to which this is an absolute limitation in the interpretation of OTL is addressed in the third challenge.

Moving Beyond Summed Scores

The decision to estimate Rasch scores, rather than simple summed scores of responses to the ordinal rating scales, was based in part on the argument above regarding statistical properties of Rasch scores versus ordinal summed scores. In addition, we intended to take advantage of advances in measurement theory and practice (Reckase, 2010), by employing a measurement model for all scale-like indices produced from TEDS-M. The Rasch model is also consistent with the model used for measure construction (EFA) and confirmation (CFA), as items are indicators of a latent trait or a domain that is larger than the simple sum of the items. Rasch analyses also provide indices of data-model fit for both items and persons. It is also consistent, although a much different model, with the summed score approach, as in the Rasch model, the total score summarizes completely a person's position or location on the variable being measured; with the additional characteristic that a comparison between two people is independent of the specific items used within a set of items indicating the same variable.

Improved Item Analysis

As an example, consider the Coherence scale (MFB15COH). In Table 2, we see the Infit and Outfit mean-squares (MNSQ) are all much smaller than 2.0 (a typical criterion for model-data fit for items). We also see consistently high pointbiserial correlations (PTBISERL CORR.)

between the items and total scores (.59 to .75), another indicator of the appropriateness of the Rasch model.

Table 2
Item Statistics Table from Winsteps for Coherence

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBISERL-EX CORR.	PTBISERL-EX EXP.	ITEM
131	53964	18747	.555	.014	1.01	1.0	1.02	1.4	.66	.67	MFB015A
132	57931	18640	-.317	.015	1.32	9.9	1.34	9.9	.59	.68	MFB015B
133	55503	18576	.237	.013	.85	-9.9	.84	-9.9	.74	.70	MFB015C
134	57277	18518	-.259	.015	.89	-9.9	.86	-9.9	.74	.70	MFB015D
135	55798	18484	.057	.014	.88	-9.9	.87	-9.9	.75	.71	MFB015E
136	57404	18481	-.273	.015	.99	-.6	.98	-2.0	.70	.70	MFB015F

With the OTL scales, the idea is not that we believe there is a "measure" or "trait" underyling the indicators of OTL, but that we can do a better job of ordering people on this scale than a simple summed total score. The person is characterized on a score that is mathematically derived from the invariance of comparisons among persons and items. Given the additional information obtained from Rasch scaling, regarding the functioning of items as an indicator of the underlying variable, this model is superior to the simple summed score. Summed scores are not exempted from the idea that they must have something in common, such as an underlying trait, and are often interpreted in ways that are only supported from the kinds of assumptions and analyses obtained from a Rasch analysis.

Measures constructed regarding the beliefs about teaching and learning mathematics are more classically aligned with the ideas of measurement and scaling, as they constitute underlying traits or constructs, from which a sample of items has been identified. However, they are similarly lists of observations of individuals using indicators (items) of a composite construct.

Assessing the Functioning of the Rating Scale

Another indicator of item quality provided by Rasch analysis is the structure of the rating scale for each item. One question is: Are the points on the rating scale ordered consistently with the overall measure (scale score)? We can observe this through examination of item characteristic curves condition on the overall measure, as in Figure 1. Here we see the four rating scale categories all represented with some nonzero probability, in order (from 1 to 4). We see this as each of the four points have an observed probability. As we move across the measure scale, the probability of endorsing the item at a higher level increases; specifically, from the lowest scores to -1.97, the highest probability is in the lowest category of 1, from -1.97 to -0.12 the probability shifts to category 2, from -0.12 to 3.76 the probability shifts to category 3, and from 3.76 to higher scores the probability shifts to category 4. Across these categories and across the measure scale, we observe order of the rating scale points.

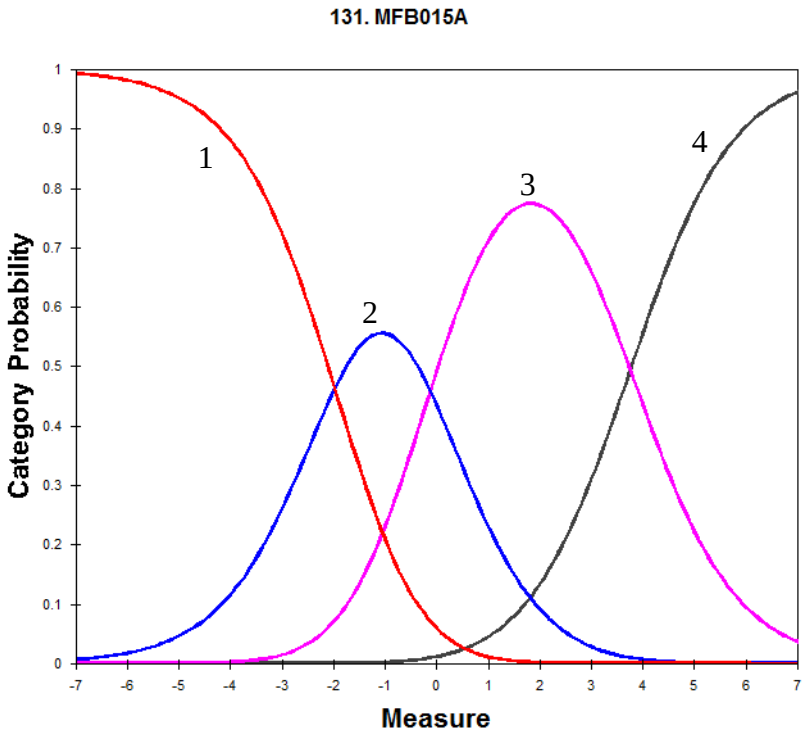

Figure 1. Item (threshold) characteristic curves for item B15A from the coherence measure.

As an example of an item with more challenging characteristics, consider item D001H from the measure of Beliefs about the nature of learning mathematics as a process of inquiry (MFD1PROC). This scale was measured with a set of items using a six-point rating scale. Notice for this item, scale points 2 and 3 (especially 3) are not observed with clear probability (Figure 2). This suggests some disorder in the rating scale points, where perhaps six points are not necessary or the six points as defined (labeled) are not functional given the demands of the item. Fortunately, for nearly all other six-point rating scale items in the beliefs measures, all six points functioned well, similar to the example in Figure 1.

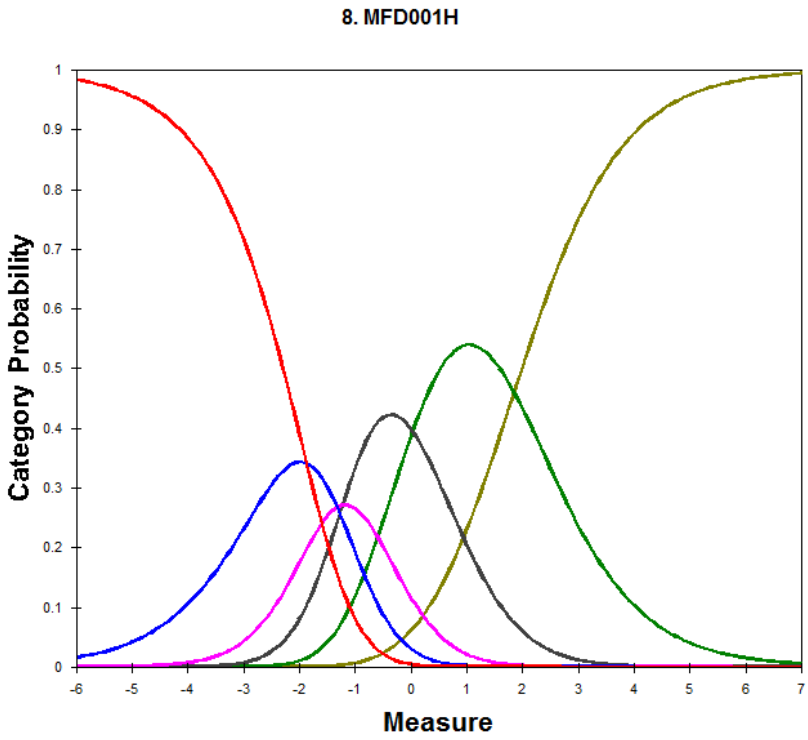

Figure 2. Item (threshold) characteristic curves for item D1H from the measure of beliefs about the nature of learning mathematics as a process of inquiry.

Scaling Measures

To complete the scaling, OTL and Beliefs indices were scaled independently, using a combined file of future primary and secondary teachers across countries. Only those cases that responded to more than 50% of the items were included in the scaling. Weights were recomputed for each OTL index, accounting for the variation in the resulting sample based on inclusion criteria (response to more than 50% of the items within a scale), since each scale was responded to by a different proportion of respondents within each country. These weights were then adjusted again so that they sum to 500 for each country for primary and secondary separately. That is, each country with primary and secondary respondents contributed 500 primary and 500 secondary units of observations to the final scaling. The weights were estimated using a simple transformation based on resulting sample size and effective sum of 500 for each population in each country. This first level of analysis with valid cases constituted the calibration sample.

The calibration values were then used to provide scores for all cases responding to more than 50% of the items. This was done to provide scores for all cases, even those excluded based on sample adjudication, allowing countries with cases not included to conduct full analyses of all cases within countries, as deemed meaningful within each country.

Building in Interpretability into Scale Scores

To facilitate improved score interpretation, scores were rescaled. Because of the one-to-one correspondence of summed scores to θ-measure in the Rasch model, we were able to rely on

the test characteristic curve (essentially, the one-to-one correspondence table between summed score and Rasch measure) to relocate final scale scores such that the scale score of 10 is associated with the midpoint of the raw score scale (the point half-way between never and often, or between disagree and agree). This provides for a common interpretable metric for OTL and Beliefs indices, such that 10 is associated with a mid-point regarding frequency, a neutral perspective regarding agreement, or a midpoint regarding the extent of preparedness (for example) for each index. As can be seen in Figure 3, the midpoint of the raw score scale (6 to 24) is 15, which is associated with a Rasch score of -0.318.

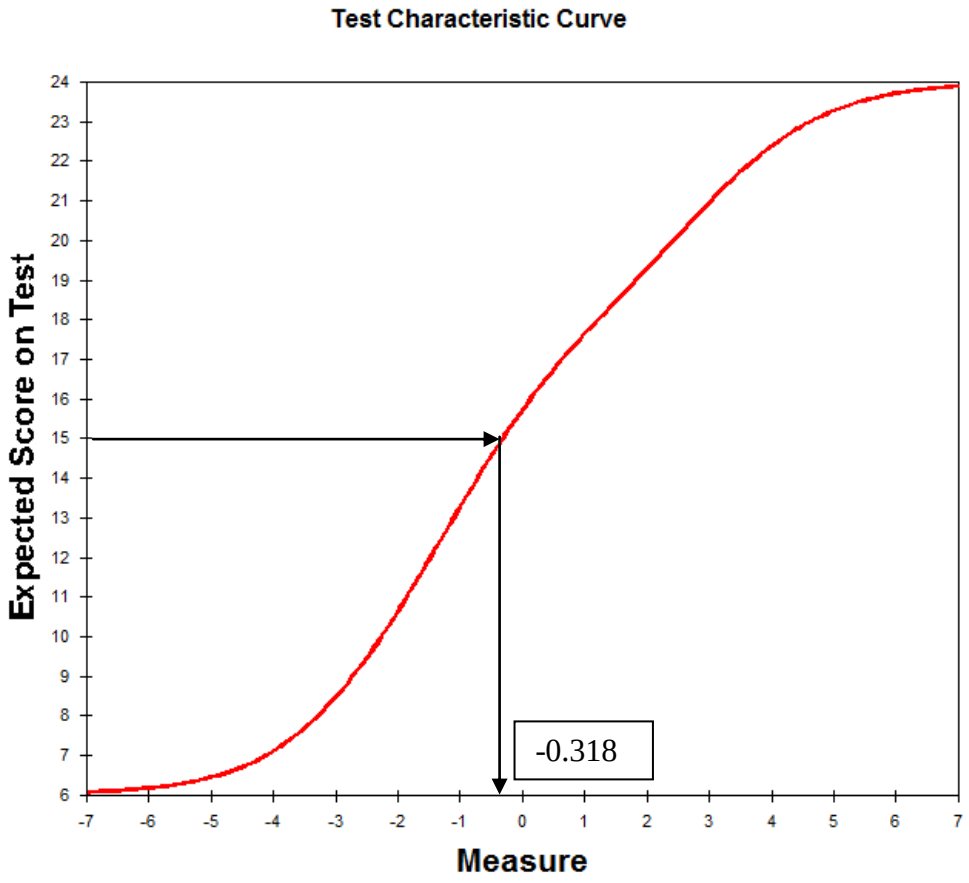

Figure 3. Test characteristic curve for Coherence (MFB15COH).

To complete the scaling, the person measures (on the logit metric) were converted to scaled scores by subtracting -0.318 from each score and adding 10:

ScaleScore = Person Rasch Measure – (Rasch score associated with middle scale point) + 10.

So for the Coherence example, MFB15COH = Rasch Measure – (-0.318) + 10, and with an original Rasch Measure of -0.318, the MFB15COH score = -0.318 – (-0.318) + 10 = 10. The Rasch Measure associated with the middle of the rating scale points is then converted to 10. This value is actually obtained from the table of summed score to Rasch score conversions (Table 3).

Table 3
Summed Score to Measure Conversion Table from Winsteps for Coherence

SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.	SCORE	MEASURE	S.E.
6	-5.416	1.862	13	-1.105	.612	20	2.426	.770
7	-4.117	1.068	14	-.724	.625	21	3.027	.787
8	-3.279	.806	**15**	**-.318**	.651	22	3.695	.862
9	-2.718	.703	16	.130	.691	23	4.617	1.102
10	-2.263	.651	17	.642	.740	24	5.966	1.880
11	-1.859	.624	18	1.220	.776			
12	-1.479	.612	19	1.828	.778			

The score scale is not designed to facilitate direct interpretation, but to capture variability in person location on a measure developed through a measurement model (Rasch) to facilitate modeling of OTL and beliefs in the study of teacher preparation in mathematics. This is a larger purpose beyond simple interpretation of these measures alone. However, to facilitate some degree of interpretability, centering the scales at 10 as a neutral or middle point provides some guide toward improved interpretability. In this way, scores greater than 10 suggest that individuals or groups report to have greater OTL (engaging in some academic content, practice, or activity occasionally or more) or agreeing with a given belief. The strength in the resulting scales is improved statistical properties to facilitate modeling of important outcomes of teacher preparation, employing OTL and beliefs on continuous scales built from items that function well as composites.

Challenge 3: Individual-Level Measures of Program-Level Characteristics

One challenge faced by the management team and discussed with National Research Coordinators is how to consider measures of OTL and beliefs as individual-level of program-level characteristics. In the conceptual framework, OTL is conceptualized as a characteristic of teacher preparation programs, as described above. Beliefs are conceptualized as learning outcomes of the preparation program, as well as knowledge. Because OTL is measured at the individual level, the question of how to score and report these measures is an appropriate one. If OTL is actually a program level characteristic and should not be considered at the individual level, scores should have been estimated for programs and not included in the individual-level database. To include these measures only at the program level would ignore variability among individuals. But if OTL is truly a program characteristic, variability among individuals might be conceived of as random error or measurement error.

In the conceptual framework, OTL is described as being "central to explaining the impact of teacher preparation on teacher learning… OTL in the study serves a number of purposes: as an explanation of differences in levels of knowledge; as an indicator of curricular variation among countries; as an aspect of fairness (e.g., appropriateness of language of test items); and as a representation of the diversity of content, both overall and for distinct groups of teachers" (p. 44). These purposes raise the issue of whether OTL can vary within programs or should be considered strictly between programs. For example, is it possible that there are "distinct groups of teachers" within a program that experience a different curriculum. This is an important

question regarding fairness in the education of American minority students, where some have challenged the opportunity to learn of children of different backgrounds within the same school. This issue can be addressed from a conceptual perspective and an empirical perspective. Finally, this is an area in need of a great deal of additional study, much of which can be supported through the use of the TEDS-M database.

Conceptual Considerations of OTL

Is opportunity to learn a question about what students have the opportunity to learn, what students attempt to learn, or what they actually learn? Researchers have described these aspects in different ways, for example, curriculum can be conceived of as intended, planned, and enacted (Kurz, Elliott, Wehby, & Smithson, 2010). We might consider another phase regarding the realized curriculum, one that a student is not only exposed to (enacted) but that a student experiences and acknowledges that experience. Most might argue that the intent in measuring OTL from student feedback is to gauge the realized curriculum. Consider the following questions from Section B of the survey:

1. Consider the following topics in university level mathematics. Please indicate whether you have ever studied each topic (with response options *studied* and *not studied*).
5. In the mathematics education courses that you have taken or are currently taking in your teacher preparation program, how frequently did you do any of the following (with response options *never, rarely, occasionally*, and *often*)?
14. To what extent do you agree or disagree with the following statements about the field experience you had in your teacher preparation program?

These questions are not asking if students had the opportunity to do various things in their programs or whether these were even options for them in their preparation. The questions explicitly ask whether students have engaged in various activities, or whether they agree or disagree with various aspects of their field experience. The question is not asking if these activities were part of the teacher preparation program per se, but whether the student engaged in these activities. So OTL in this context might be considered more behavioral and a function of future teacher engagement with the program, as an index of the realized curriculum. Clearly, if the program did not offer such opportunities, students would not engage in them, suggesting limited OTL. However, if students do not engage in various activities and those activities are provided, that is students have the opportunity to participate, this would not be captured by asking questions about participation.

Is it possible for individuals to experience different levels of OTL within the same teacher preparation program? To the extent that individual students can select courses and other program opportunities to complete their program in one or more areas, OTL might be an individual level characteristic, or at least OTL, as defined by individuals, may vary among individuals within a program. However, just because a student fails to take advantage of one or more elements of a program offering does not necessarily alter the "opportunity" to learn, only the realized learning or resulting learning.

Empirical Considerations of OTL

We can also answer one question empirically: Does OTL as reported by future teachers vary within teacher education program? On one hand, one might argue that if OTL is a program characteristic, in terms of an enacted curriculum, all students within the program might report the same information. However, if OTL is an individual-level characteristic, regarding the realized curriculum, then students within a teacher preparation program will report very different levels of experiences. Although these two scenarios are quite different, perhaps two extremes, the more likely scenario is probably somewhere in the middle.

We can estimate variability among future teachers within program through a multilevel analysis of OTL scores. Because variability among programs depends a great deal on country, country will be included in the modeling of variability. HLM was used to partition variability of all OTL measures between individual Future Teachers, teacher preparation program or institution, and country. Future Teachers can be considered members of a particular higher education institution or specific teacher preparation program within institutions; so in terms of OTL, variability within both were examined. Finally, only nine countries had sufficient numbers of future teachers and institutions or programs to sustain the HLM analyses, so country level results are tentative (a sample of 9 is small); over 4000 future teachers were represented by over 120 programs. For example, Botswana has four institutions and Singapore has one included in the Primary education sample.

Table 4
Variance Components for Three-Level HLM Models of OTL Measures

	Supervising teacher feedback quality		Math ed pedagogy class participation		Program coherence	
Model Levels	Variance	Proportion of total	Variance	Proportion of total	Variance	Proportion of total
Future teacher	5.67	.86	2.15	.68	4.48	.76
Program	0.08	.01	0.56	.17	0.18	.03
Country	0.85	.13	0.47	.15	1.24	.21
Total	6.61		3.18		5.89	

In all cases, the level of variance at the program and country level was significantly different than zero ($p<.001$); except supervising teacher feedback quality ($p=.002$). However, as can be seen by the examples in Table 4, the magnitude of variance due to program (proportion of total) was relatively small. These examples suggest that there is much more variability within program (within country) than between program or between country. Similar results (nearly identical) were found when modeling institution in place of program.

Examining across all OTL measures at the primary education level (Table 5), the smallest proportion of variance accounted for by program was in the school-based experience measure of supervising teacher feedback quality (MFB14STF), which had a program proportion of variation of .01; the largest proportion of variance accounted for by program occurred with the measure of school-level mathematics of Numbers, Measurement, and Geometry (.23) followed by engaging in class participation in mathematics education pedagogy courses (.17).

Table 5
Proportion of OTL Measure Variance Accounted for by Future Teachers and Programs

Category	Index/measure	Future teachers	Program
Mathematics content			
Tertiary level mathematics	Geometry	.76	.12
	Discrete Structures & Logic	.64	.16
	Continuity & Functions	.64	.10
	Probability & Statistics	.65	.14
School Level Math	Numbers Measurement Geometry	.69	.23
	Functions Probability Calculus	.75	.10
Mathematics education pedagogy	Foundations of math ed pedagogy	.78	.04
	Principles of instruction and curriculum	.79	.15
	Engage in class participation	.68	.17
	Engage in class readings	.79	.10
	Engage in solving problems in class	.69	.16
	Instructional practice and techniques	.70	.13
	Instructional planning techniques	.86	.07
	Uses of assessment	.79	.08
	Assessment practices and techniques	.83	.11
Education pedagogy	Social sciences (history, philosophy, sociology)	.85	.05
	Application of general education pedagogy	.91	.04
	Teaching for diversity	.67	.08
	Teaching for reflection on practice	.83	.03
	Teaching for improving practice	.87	.05
School-based experience	Connecting classroom learning to practice	.83	.06
	Supervising teacher reinforcement of university goals for practicum	.92	.04
	Supervising teacher feedback quality	.86	.01
The nature of mathematics	Mathematics is a set of rules and procedures	.76	.04
	Mathematics is a process of inquiry	.77	.04
Learning mathematics	Mathematics is learned through teacher direction	.65	.03
	Mathematics is learned through active learning	.85	.02
Mathematics achievement	Mathematics is a fixed ability	.70	.04
The program as a whole	Program coherence	.76	.03
	Instructional quality	.76	.14
	Preparedness for teaching mathematics	.73	.09

From this analysis, some trends can be observed. First, all OTL and Beliefs measures at the primary education level vary significantly more at the future teacher level and much less so at the program level. For most OTL measures, the proportion of variance accounted for by programs is less than 10%. Mathematics content (topics studied) tend to be more consistent within program, where programs account for 10 to 20% of the variance (future teachers account

for less than 70% for most measures). Similarly, there is a tendency for the proportion of variance accounted for by programs to be larger with measures of mathematics education pedagogy OTL (mostly over 10%, excluding foundations, instructional planning techniques, and uses of assessment). All measures of education pedagogy and school-based experience OTL are less than 10%. As expected, the proportion of variance due to programs for beliefs measures was very small (all less than 5%).

There is a great deal of variability at the future teacher level within programs within countries. Future teachers are no reporting consistent experiences with OTL, suggesting a large discrepancy between the enacted curriculum and the realized curriculum. The argument here is that researchers analyzing data within a given country should examine the degree to which OTL measures are explained by program or institution versus future teachers. They should consider the option of including measures of OTL at both the future teacher and the program level, to account for the enacted curriculum (at the program level) and the realized curriculum (at the future teacher level). This distinction might very well be a function of country, and may be different for programs preparing secondary education future teachers of mathematics.

Discussion & Future Directions

The process of gathering information on the functioning of the TEDS-M measures according to the context and outcomes of teacher education internationally in some cases challenged the theory behind the method (such as whether the scale measure the same thing in every country). For example, using a confirmatory factor analysis approach to assess measurement invariance across countries was not possible because of some nuances in response patterns in some countries. The challenges and successes in building relevant measures are presented. However, many questions and issues remain in all three areas of measurement invariance, Rasch scale stability and appropriateness, and the nature of OTL as individual versus program characteristics. A sample of possible questions and issues for continued research on measures of OTL and beliefs is presented here.

Measurement as a field embodies a strong pragmatic perspective when it comes to application of its methods, perhaps more so than educational statistics. Although modern measurement theories are based on models requiring strong assumptions, many of which are very difficult to achieve in practice (e.g., measurement invariance), measurement is fueled by the need to learn about individual differences to inform decision making. The need to make practical decisions and move forward requires compromise between theory, modeling, and application. Much of what drives this pragmatism is our hope to learn from our data. However, validity-related evidence to support the intended inferences and uses of measures must be gathered to a degree that allows us to learn.

Questions and Issues for Continued Research on Measuring OTL and Beliefs

1. Continue to explore issues related to measurement invariance.
    a. Are there subsets of countries (one or more) where invariance is more difficult to achieve? What are the characteristics of their programs?
    b. Are there items within measures that are functioning inconsistently across countries? A model of partial measurement invariance should be investigated, identifying items within measure that are invariant across countries.

    c. A systematic investigation of sources of measurement variance will allow researchers to make sense of the meaningfulness of measures across countries.
    d. Are there other groupings for which measurement invariance should be investigated, such as gender, or native-language background of participants within country, or primary education versus secondary education future teachers, or others?
2. Investigate Rasch scale stability and meaningfulness.
    a. In part, the stability of Rasch scaling is associated with measurement invariance properties. Do Rasch estimates vary as a function of country? DIF analysis could provide useful information in this regard.
    b. Across primary and secondary education levels, the weighted SDs of measures scaled through Rasch range from 0.9 to 2.7. Since the model determined the SD for each measure, what are the characteristics of measures with larger variability?
3. Assess the nature of OTL as an individual-level measure
    a. The extent to which OTL is a program characteristic should be investigated within country.
    b. Are there individual-level characteristics that lead to greater variability in reported OTL within programs?
    c. Are there program characteristics that lead to greater variability in reported OTL?

The development of sound measures is essential to the quality of scientific studies of teacher education. By developing valid and reliable measures TEDS-M has taken an important step in this direction.

References

AERA, APA, & NCME. (1999). *Standards for educational psychological testing*. Washington DC: AERA.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.

Graham, J.M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930-944.

Haertel, E.H. (2006). Reliability. In R.L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 111-153). Westport, CT: Praeger Publishers.

Harwell, M.R., & Gatti, G.G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105-131.

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Kurz, A., Elliott, S.N., Wehby, J.H., & Smithson, J.L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *Journal of Special Education, 44*(3), 131-145.

Linacre, J.M. (2009). Winsteps Rasch measurement computer program (Version 3) [Computer software]. Beaverton, OR: Winsteps.com.

Muthén, L.K., & Muthén, B.O. (2007). Mplus: Statistical analysis with latent variables (Version 5) [Computer software]. Los Angeles, CA: Authors.

Reckase, M.D. (2010). NCME 2009 Presidential address: "What I think I know." *Educational Measurement: Issues and Practice, 29*(3), 3-7.

Tatto, M.T., Schwille, J., Senk, S.L., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., & Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries.* Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement.

Tatto, M.T., Schwille, J., Senk, S., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher Education Study in Mathematics (TEDS-M), conceptual framework*. East Lansing, MI: Michigan State University, TEDS-M International Study Center.

Tatto, M.T., Senk, S.L., Bankov, K., Rodriguez, M., & Peck, R. (2011). *TEDS-M 2008 Assessment Frameworks: Measuring future primary and secondary teachers mathematics and mathematics pedagogy knowledge*. East Lansing, MI: Michigan State University, TEDS-M International Study Center.

Tatto, M.T., & Senk, S. (2011). The mathematics education of future primary and secondary teachers: methods and findings from the Teacher Education and Development Study in Mathematics. *Journal of Teacher Education, 62*(2), 121-137.

Vandenberg, R.J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70.