Conceptualizing and Measuring Opportunities to Learn and the Contexts of Teaching

Michael C. Rodriguez, University of Minnesota
Teresa Tatto, Michigan State University

April 20, 2015

Paper presented in the symposium

Toward a Fair Evaluation of Teachers: Methodological Challenges in a Cross-National Study of Mathematics Teachers
(Organizer, Maria Teresa Tatto)

at the annual meeting of the
American Educational Research Association, Chicago, IL.

This research is sponsored by a grant to Michigan State University from the National Science Foundation [DRL 0910001]. The development work has been done in collaboration with the University of Minnesota, University of Nebraska Lincoln, and the World Bank. International collaborators from more than fifteen countries have been supported by their countries research foundations, their own governments, the World Bank, and the Inter-American Development Bank.

INTRODUCTION

The FirstMath study is an international collaboration, following up on the highly successful Teacher Education and Development Study in Mathematics (TEDS-M), investigating the national and local contexts of teacher preparation, opportunities to learn and teacher preparation institutional contexts, and the mathematics knowledge and beliefs about teaching and learning mathematics of future teachers in the last year of teacher preparation (see Tatto et al., 2012).

The FirstMath conceptual framework expands on the TEDS-M framework, at is investigates the contexts, knowledge, and practices of novice teachers, mathematics teachers in the first five years of teaching. This model includes novice teacher characteristics, their beliefs about teaching and learning, their teaching practices, national policy and local contexts, student characteristics, and the mathematics teaching knowledge of novice teachers as well as the mathematics knowledge of their students. This framework is informed through a series of instruments to be administered internationally, including:

- Teacher Knowledge Test (separate primary and secondary level tests)
- Teacher Questionnaire
- Teacher/Classroom Observations
- Curriculum Information
- National & Local Policy information
- Student Knowledge Test & Background Questionnaire

The focus of this paper is in the novice teacher questionnaire (NTQ) and the components of that instrument addressing beliefs and practices. Although the beliefs have been fully developed for future teachers, they have only been field tested with novice teachers. The opportunities to learn (OTL) components are slightly modified as they are both retrospective (in terms of preparation for teaching) and current (in terms of learning through practice and professional development). The components addressing teaching practices are being newly developed, although based on the research literature regarding effective and quality mathematics teaching practices.

The design of the novice teacher questionnaire spanned several phases, including an early item pilot, a formal field trial, and final instrument review and design. A formal process was developed to ensure coherence and consistency in item format and presentation for all items in all of the instruments. This paper describes our process and presents three key challenges to be addressed in realizing the goals of FirstMath.

Theoretical framework

This paper is primarily methodological, describing the methods of conceptualizing and developing measures of OTL and the contexts of learning. These methodologies include 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, 2012) and Winsteps to employ the Rasch scaling model (Linacre, 2012).

The survey items come from a variety of sources, including the Teacher Education and Development Study (TEDS-M) and the Third International Mathematics and Science Study (TIMSS) (Provasnik, et al., 2012), and focus on the key elements of mathematical knowledge for teaching (Ball, Thames, & Phelps, 2008). These sources provide a deep theoretical framework to the conceptualization of the FirstMath study, much of which relies on the TEDS-M conceptual framework (Tatto et al., 2008).

Regarding the methodological approaches employed in the design, analysis, and reporting of FirstMath results, a contemporary validity framework is the essential basis.

A Validity Framework and Three Related Challenges

A core element of any study is measurement quality, which is required to support inferences from any measure. Validity is the 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, 2014). "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, NCME, 2014, p. 11). 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, as they are relevant to intended score interpretations and uses.

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 FirstMath results? The primary inferences intended from the measures of beliefs, OTL, and contexts of teaching focus on the content of those measures, particularly regarding OTL and contexts. The very idea of OTL presents a challenge to measurement (more on this in a moment).

Another core inference regards the appropriateness, meaningfulness, and usefulness 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 countries, and relations to other variables (as this is the key function of the conceptual framework, as beliefs, OTL, and teaching contexts are viewed to be critical moderators and in some respects outcomes regarding teacher effectiveness.

Content-related validity evidence is found in the FirstMath conceptual framework, the assessment frameworks, all of which will be documented in the technical manual. Additional evidence is found directly in the items themselves. The specific items used to create each measure will be available in the public database (following operational completion of the main study) and others will able to evaluate the functioning of these items as indicators of each measure. However, the provision of this evidence is the responsibility of the measure developer.

Evidence of relations to other variables will be the primary focus of future analyses, as this is the function of the FirstMath study. Many researchers will use the TEDS-M database to evaluate the associations among beliefs, OTL, teaching practices and teaching contexts, knowledge measures, and many other background variables, and characteristics of students as well as student knowledge. The extent to which measures of beliefs, OTL, and teaching contexts function as intended in the conceptual framework provides additional validity-related evidence.

These sources of validity evidence and the entire validation process present the first major challenge to the project (Challenge 1). In particular, because of the international comparative nature of the intended inferences, evidence regarding measurement invariance is important to provide (Challenge 2). Finally, the development of scales that are appropriate, meaningful, and useful is another major challenge, where we have some experience from TEDS-M (Challenge 3).

METHODS

In all stages of the study, item development analyses were conducted, including exploratory factor analysis, and correlations among and between items of similar and different constructs. The items functioned exceptionally well compared to prior experience with pre-existing items. Some items underwent revision based on item pilot reviews and we expect additional item refinement based on Field Trial results. A standard set of item-writing guidelines were adopted to assure consistency and coherence in all measurement aspects.

Developing FirstMath Measures

As stated in the background discussion above, several approaches were employed in the iterative development stages, all focusing on the extent to which concepts and items were relevance to novice teachers of Mathematics across countries. In addition, we heavily relied on quality evidence from TEDS-M. The pilot study was conducted in several countries, resulting in complete responses from 83 novice teachers, followed with a more comprehensive field trial, employing a structured sampling design across 12 countries, resulting in 380 novice teachers.

The process of item review was comprehensively completed across the pilot results and underway with field trial data, through a collaborative process with national research coordinators and experts associated with each country. These procedures also deepened the team's understanding of each area of measurement.

Instrumentation

The full set of instruments included the following:

- Teacher Mathematics Teaching Knowledge Test
- Teacher Questionnaire (NTQ)
- Teacher/Classroom Observations
- Curriculum Information
- National & Local Policy information
- Student Knowledge Test and Background Questionnaire

The focus of this paper is on the NTQ, which contained the following major and subsections

I. Beliefs about Teaching and Learning
   A. Math is a set of rules and procedures
   B. Math learning should be teacher directed
   C. Math is a fixed ability
   D. Level of preparedness to teach mathematics

II. Opportunities to learn in Teacher Preparation
   A. University level mathematics
   B. General and Mathematical Pedagogy
   C. Teaching for diversity
   D. Assessment practices
   E. Reflecting on practice
   F. Improving practice

III. School Curriculum
   A. Topics prepared to teach
   B. Topics covered this year

IV. The FirstMath Class
   A. Class description, numbers of students
   B. Instructional practices
   C. Homework practices
   D. Assessment practices
   E. Challenges faced in the classroom

V. Local Context
   A. Perceptions of working in current school
   B. Interactions with teachers

VI. Becoming a Teacher
   A. Route to teaching
   B. Level of mathematics studied
   C. Licensure/certificate status
   D. Professional development participation
   E. Other careers
   F. Reasons for becoming a teacher
   G. Future as a teacher

VII. General Background
   A. Age, gender
   B. Family socio-economic status, resources
   C. Family education levels
   D. Languages spoken
   E. Prior educational achievement

Major Challenges

As introduced above, there are three major challenges facing the measurement tasks associated with achieving the goals of FirstMath. These include:

1. Validation
2. Measurement Invariance
3. Scaling and Scale Meaning

Challenge 1: Validation

To support the rigorous development and validation of each measure, we followed a measurement design model that is best characterized by Wilson (2005), in the translation from the construct to item responses through the outcome space and a measurement model, supporting inferences back to the construct.

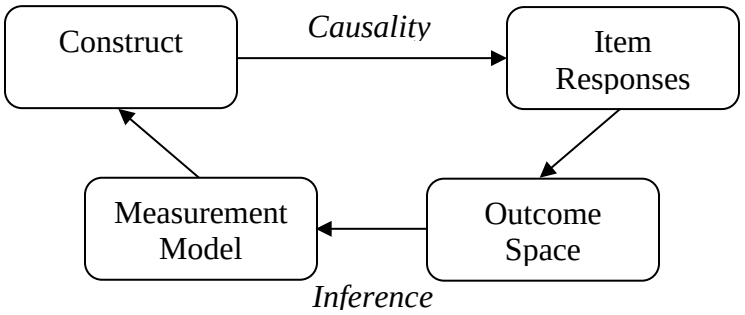


Figure 1. Wilson's (2005) model of measurement design.

From Figure 1, we see four interacting components of measure design that also provide stages at which validity evidence can be gathered and documented.

- The construct (trait level) as manifested in an individual ***causes*** the observed item responses.
- Item responses are observed realizations of the construct.
- The outcome space includes the scoring routines, response options, those aspects of responses we value – how we score observations, ratings, responses.
- The measurement model specifies how we relate scores to constructs; in our case, the Rasch measurement model.

Specifically, we base the content of the measures on prior research, theory, and practice – particularly given the international contexts of our work and review of researchers and practitioners in partnering countries. These items were developed through an iterative process, including gathering items currently in use, requesting items from international partners, and conducting a comprehensive review of the relevance and appropriateness of each item. Through the 2012 Pilot, item responses were evaluated and response options were evaluated to assess the extent to which they mapped onto expected levels of functioning. In all, over 500 questions were piloted in the NTQ. Items that resulted in low levels of variability and inconsistent relations to

other items were eliminated. A much reduced set was again reviewed internationally and administered in the 2014 Field Trial, including 380 questions. Field Trial data are being evaluated to investigate the extent to which item responses function in the measurement model – the Rasch partial-credit model.

The work is grounded in a strong validity framework, where validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014). From this perspective, one of building an argument to support the intended interpretations and uses (Kane, 2013), we specified a set of intended inferences. Specifically, the measures:

- include **core content of OTL** for teacher preparation in Mathematics;
- capture **core beliefs** about teaching and learning mathematics;
- include **core practices** important for effective mathematics teaching;
- are appropriate, meaningful, and useful across **international contexts.**

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 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 FirstMath, given the successful scaling of similar measures in 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, since in the Rasch model, the total score summarizes completely a person's position or location on the variable being measured.

Challenge 2: CFA & Measurement Invariance

Measurement invariance involves a class of analyses addressing the invariant functioning of a measure across important groups of participants. Our primary concern is the invariance of the beliefs, OTL, and teaching contexts measures across countries. To accomplish the assessment of measurement invariance, in part through comprehensive scaling and assessment of data fit, at least the following tasks will be completed based on operational data prior to final scaling and reporting.

- Multiple Group Confirmatory Factor Analysis (MCFA). This provides for a test of the fit of a given factor structure in each country.
- DIF analysis, to assess fit of items within each measure across countries.
- Accommodation of missing data, and setting decision rules for the magnitude of missingness allowed to score the item responses for a given individual.
- Estimation appropriate with non-normal discrete factor indicators, accommodating the rating scale data appropriate (Mplus employs a probit-regression model to complete CFA with ordinal data).

Challenge 3: Scaling and Scale Meaning

Finally, we have an opportunity to create a scale that conveys some useful information. The Rasch model, as with most IRT models, creates a scale that is based on a logistic metric, centered at zero (with Rasch, the average item location is used to define the zero on the scale), generally with a standard deviation about 1.0, resulting in a scale that generally ranges from -4 to +4. This scale doesn't have any inherent meaning, since it is basically arbitrary.

The resulting scale does have some nice properties. The Rasch model, as with all IRT models, places items and persons on the same scale – providing a tool for score interpretation. In TEDS-M, concurrent calibration was used to scale beliefs and OTL measures across primary and secondary levels of future teachers, providing for a consistent scale and interpretation across levels for which teachers are being prepared to teach.

In addition, a number of beliefs and OTL measures are the same or related to measures used in TEDS-M. Because of the prior successful Rasch scaling of many of these measures, we have additional potential to support score interpretation in FirstMath. For instance, we can consider the following two questions:

- Can we place (equate/link) some measures on the TEDS-M scale?
- Can we support inferences about differences between Future Teachers and Novice Teachers?

The method of centering and transforming the Rasch scores, as was done in the TEDS-M study, will be used in the FirstMath study to support score interpretation. This is explained more fully below.

RESULTS

The process of gathering information on the functioning of the study's measures according to the context and outcomes of teaching internationally in some cases challenged the theory behind the method (such as whether a scale measures the same construct 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. Although the Field Trial data are limited in country-specific sample sizes, the overall data set is sufficient to begin examining measure quality overall.

To address the two challenges (Challenges 1 and 2) described above, initial examples are provided here to indicate their intent. In part, the examples here are based on the 2014 Field Trial, including participation from the following countries:

- Bulgaria
- Chile
- England
- Guatemala
- Guyana
- Honduras
- Mexico
- Peru
- Philippines
- Slovakia
- Turkey
- USA

These data included 380 novice teachers, 870 primary pupils, and 850 secondary pupils. For the purposes here, only the data from the 380 NTQs are included.

Challenge 1: Rating Scale Structure Supporting CFA and Measurement Invariance

The challenges in evaluating measurement invariance are numerous and well documented in the literature (Brown, 2014). At a basic level, the evaluation of measurement invariance begins with a test of the factor structure fit across countries by testing the factor loadings of items and their invariance across countries.

With international participants, one issue is the presence of response sets – or tendencies of some, perhaps due to cultural influences, to not select the extreme values. Or, perhaps in some countries, there is more uniform beliefs or contexts that result in less variability in responses, so that the range of options selected is limited.

Consider the measure of School Conditions. The responses by country for two items are listed below in Tables 1, 2, and 3. Here we see that there are small sample sizes. But imagine that the samples are sufficient for interpretive processes, and observe the extent to which some response options have zero frequencies.

In Table 1, for example, only two countries (J, K) have observed responses in the Very Low category. One country (J) has zero observed responses in the Very High category. Three countries (C, H, I) have zero observed responses in both the Low and Very Low categories. What this does is change the rating scale structure of this item across countries.

For nine countries, this item has four categories; for three countries this item has three categories. This result will result in a failure to claim invariance (since the scale structure changes across countries, its functioning will be variant).

Table 1.
Characterizing School Conditions Regarding Teachers' Job Satisfaction

Country	*Very high*	*High*	*Moderate*	*Low*	*Very low*	*Total*
A	6	6	3	1	0	16
B	6	13	4	1	0	24
C	2	3	2	0	0	7
D	4	10	21	2	0	37
E	8	14	17	3	0	42
F	1	25	31	4	0	61
G	1	10	6	1	0	18
H	3	22	15	0	0	40
I	2	12	8	0	0	22
J	0	4	13	12	1	30
K	5	15	12	0	2	34
L	1	1	5	1	0	8
Total	39	135	137	25	3	339

In Table 2, eight countries produced 5-point scales, three countries (A, F, I) produced 4-point scales, and one country (C) produced a 3-point scale.

Table 2.
Characterizing School Conditions Regarding Parental Support for Student Achievement

Country	*Very high*	*High*	*Moderate*	*Low*	*Very low*	*Total*
A	4	4	4	5	0	17
B	2	6	6	7	3	24
C	0	1	0	5	1	7
D	4	5	12	13	2	36
E	1	4	8	17	11	41
F	0	8	17	18	18	61
G	1	3	6	7	1	18
H	3	12	12	10	2	39
I	2	5	12	3	0	22
J	1	1	4	11	13	30
K	6	5	11	8	4	34
L	1	1	4	1	1	8
Total	25	55	96	105	56	337

In Table 3, eight countries produced 5-point scales and four countries produced 4-point scales.

Table 3.
Characterizing School Conditions Regarding Parental Involvement

Country	*Very high*	*High*	*Moderate*	*Low*	*Very low*	*Total*
A	3	4	5	4	1	17
B	1	6	8	5	4	24
C	0	1	1	3	2	7
D	1	7	10	11	8	37
E	0	3	11	18	10	42
F	1	4	17	19	20	61
G	2	2	7	3	3	17
H	1	12	16	8	3	40
I	1	6	7	8	0	22
J	2	0	2	14	12	30
K	6	3	10	8	7	34
L	1	1	3	2	1	8
Total	19	49	97	103	71	339

These results help us interpret findings of non-invariance and allow us to evaluate the appropriateness of items. However, the formal evaluation of measurement invariance via the factor structure will be limited because of these structural differences in item response scales.

Other Approaches to the Evaluation of Measurement Invariance. Another important tool for evaluating measurement invariance is differential item functioning (DIF). This is a way to evaluate item functioning across important groups of participants. DIF analyses will be completed with the final data to support score interpretation across countries and across other participant characteristics, such as gender – which is an important characteristic of novice teachers. There are still significant gender differences in the entry and retention of mathematics teachers, which also varies across primary and secondary levels of mathematics. In the TEDS-M study, Albano and Rodriguez (2013) evaluated DIF by gender and employed a multilevel item response model to use OTL as an explanatory variable to understand gender DIF.

Challenge 2: Reporting & Interpretation

As mentioned in the discussion of the measurement model, the Rasch model results in an arbitrary scale, somewhat centered at zero with a range generally from -4 to +4. This limits direct interpretation of scores. However, the Rasch scale can be centered and transformed to provide at least one (if not more) points that convey response information. For example, in many state testing programs using IRT scoring methods, scales are centered so that the proficient score point

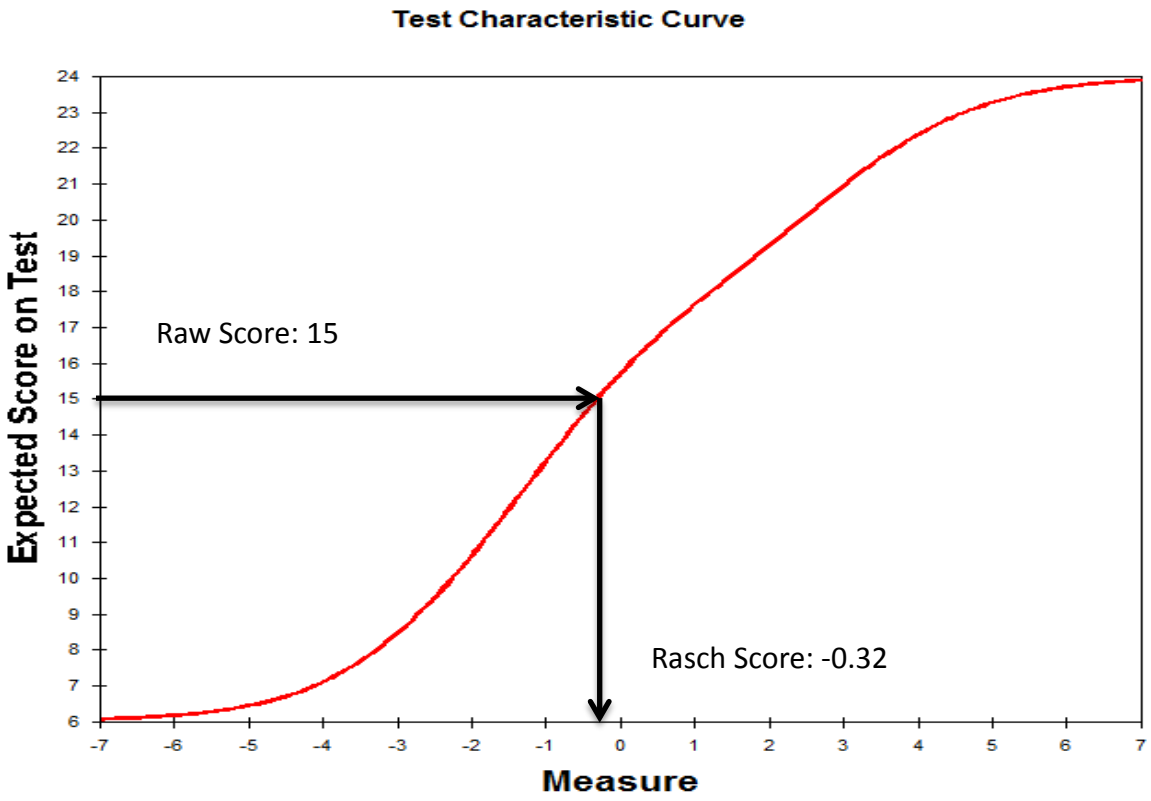
is associated with a specific value, such as 50 (e.g., Minnesota NCLB tests are all scaled so that scores of 50 are at the proficient cut point).

Some measures are naturally interpretable and do not need scaling or transformation. For example, many OTL scales, such as “Topics Studied” are simple counts of the number of topics reportedly studied by the participant. These are clearly interpretable and range from zero to the total number of topics available in a given category.

Other OTL, beliefs, and school context questions are scalable and tend to be related to perceptions (e.g., agree to disagree) or frequencies (e.g., never or almost never to daily). These measures are subjected to Rasch scaling and so require centering and transformation to support interpretation.

Based on the methods used in the TEDS-M study for some OTL and all beliefs measures, the measures are first scaled with the Rasch partial credit model. The resulting Rasch scores are then centered and transformed so that all scales are centered at 10. This is done in two steps.

First, the midpoint of the rating scale is identified on the raw score metric. For example, if the scale includes six items on a 4-point scale (1 to 4), the raw score could range from 6 (all six items are rated a value of 1) to 24 (all six items are rated a value of 4). The midpoint of this scale is 15 ([6+24]/2). Using the Test Characteristics Curve (which associates raw scores to Rasch scores), the associated Rasch score can be obtained. An example is provided in Figure 2, where the midpoint, a raw score of 15, is associated with a Rasch score of -0.32. This Rasch score is used to center the Rasch score scale around that point, so that zero is now the midpoint of the response scale (e.g., half way between agree to disagree for perception items).



Consider this fictitious example of a measure of Teaching Context, where the midpoint of the scale is found at a Rasch score of -0.32. To center the scale at this point, we simply subtract this value from each Rasch score:

[Teaching Context] = $[Rasch\ Score]_i$ – (*Midpoint Rasch Score*)

If the *Midpoint* = -0.32 and participant *i* has a Rasch score of -0.32, then $[-0.32]_i$ – (-0.32) = 0.

But with this centering, many scores remain negative and one might want to report positive scores. The scores can be transformed to be positive by simply adding a positive value to each score. In TEDS-M we recentered scores to 10 by adding 10 to each score. This was done to avoid misinterpretation of these scores as standardized values, since the mean would be zero, although the standard deviation is not necessarily one. The final transformation for each Rasch score is then:

[Teaching Context] = $[Rasch\ Score]_i$ – (*Midpoint Rasch Score*) + 10

In this way, each measure is centered at 10 so that 10 indicates the midpoint of the rating scale. Then we can estimate means and standard deviations and pursue statistical modeling with these centered and transformed scores.

The means of each belief, OTL, and school context measure are defined by location of the population. Some means will be below 10 (when the average response is in the negative region of the response scale (e.g., disagree) and some will be above 10 (when the average response is in the positive region of the response scale (e.g., agree). Although the means are defined by the location of the population, the standard deviation is set by the model – indicating the amount of ability required to produce a specific probability or likelihood of moving up one logit on the Rasch scale. The standard deviations will naturally vary across measures, based on the variability of the population in their responses to each measure.

For example, a sample of measures are reported from the TEDS-M final results. Notice that some measures (Assessment Uses, Teaching for Diversity, and Active Learning) are well above the midpoint or neutral point of 10. In the case of Active Learning, the mean of 12.0 is more than one standard deviation (1.3) above the neutral point of 10. Since these means are greater than 10, we can interpret the results as indicating future teachers had more opportunities to learn about Assessment Uses (11.0) and Teaching for Diversity (10.4) than not, and that future teachers tend to endorse the belief that it's best to learn mathematics through active learning (12.0).

In other cases, the measures are below the scale midpoint, including Teacher Direction (9.3) and Achievement (9.5) as Fixed Ability, indicating that future teachers are less likely to endorse these beliefs (that learning should be teacher directed or mathematics achievement is a function of fixed ability).

Table 4.
Descriptive Statistics of TEDS-M Scaled Measures

Primary FTs	***M***	***SD***	***Min***	***Max***
Assessment Uses (OTL)	11.0	2.5	4.7	15.3
Teaching for Diversity (OTL)	10.4	2.0	5.1	15.1
Teacher Direction (Belief)	9.3	0.9	5.0	14.8
Active Learning (Belief)	12.0	1.3	6.2	15.7
Achievement as Fixed Ability (Belief)	9.5	1.0	5.1	15.1

In this way, we are able to provide some interpretability built into the scaling metric, without losing the relative position of scores across participants and countries, and since we have not altered the variability in any way, we retain all information to obtain associations with other variables as with untransformed scores.

CONCLUSIONS & FUTURE PLANS

The development of sound measures is essential to the quality of scientific studies of teaching. Understanding the beliefs and opportunities to learn various mathematics, curricular and pedagogical topics in an international context will allow policymakers to have a research base on which to improve teacher induction. Measuring relevant beliefs about teaching and learning and similarly measuring the contexts of teaching will provide a strong basis for investigating their important roles in teaching and learning. By developing measures with strong valid evidence, this study has taken an important step in this direction.

The Field Trial data will be examined closely to evaluate the functioning of each study instrument. The other methods reported in this symposium were successfully implemented in the 2014 Field Trial and produced useful information for finalizing plans for the main operational study, including sampling methods (Mark Reckase, Teresa Tatto, and Michael Rodriguez), teacher knowledge tests (Kiril Bankov and Michael Rodriguez), and the observation tools (Wendy Smith and Teresa Tatto). Together, we have assembled a rigorously developed set of tools to investigate the OTL, beliefs, practices, contexts, and knowledge of novice teachers of mathematics.

REFERENCES

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington DC: AERA.

Albano, A.D., & Rodriguez, M.C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational and Psychological Measurement, 73*(5), 836-856.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Brown, T.A. (2014). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review,* 42(4), 448-457.

Linacre, J. M. (2012). *Winsteps* (Version 3.74) [Computer software]. Chicago, IL: Winsteps.com.

Muthén, L. K., & Muthén, B. O. (2012). *MPlus* (Version 7) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., and Jenkins, F. (2012). *Highlights From TIMSS 2011: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2013-009 Revised). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

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

Tatto, M.T. (Ed.). (2013). *The Teacher Education and Development Study in Mathematics (TEDS-M): Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries. Technical report*. Amsterdam: IEA. Retrieved from http://www.iea.nl/teds-m.html

Tatto, M.T., Schwille, J., Senk, S.L., Ingvarson, L., Peck, R., & Rowley, G. (2008). *Teacher education and development study in mathematics (TEDS-M).* Amsterdam: IEA. Retrieved from http://www.iea.nl/teds-m.html

Tatto, M.T., Schwille, J., Senk, S.L., Ingvarson, L., Rowley, G., Peck, R., Bankov, KI., Rodriguez, M., & Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries*. Amsterdam: IEA. Retrieved from http://www.iea.nl/teds-m.html

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.