Chapter 9

Measured Mathematics

Jeremy Kilpatrick
University of Georgia

Research in mathematics education involves phenomena, constructs, and instruments. The phenomena concern the teaching, learning, and doing of mathematics, each of which has multiple aspects. The constructs touch on selected aspects of those phenomena. The instruments are designed to measure attributes of the constructs. Mathematics educators concern themselves with making sure that the constructs reflect the phenomena, and psychometricians concern themselves with making sure that the instruments reflect the constructs. The preceding chapters in the monograph have brought mathematics educators and psychometricians together to consider how each field might profit from the other. In this commentary, I respond to those chapters from the perspective of a mathematics educator.

From a distance, proficiency in mathematics can look like a one-dimensional phenomenon, something that people possess to a greater or lesser degree. For example, in a study from Harvard's Program on Education Policy and Governance and the journal *Education Next* (Hanushek, Peterson, & Woessmann, 2010), which found that the United States ranked behind most of its industrialized competitors in mathematics performance, researchers compared the performance of high achievers not only across countries but also across the 50 U.S. states and 10 urban districts. Most states and cities ranked closer to developing countries than to developed countries. To achieve the rankings, the researchers compared the percentages of U.S. students in the states and districts who performed at an advanced level in mathematics on the 2005 National Assessment of Educational Progress (NAEP) with estimated percentages of students in other countries who would have reached that same level had they taken the NAEP 2005 mathematics assessment. The rankings required that NAEP and Program for International Student Assessment (PISA) mathematics scores use the same scale, so the researchers "assumed that both NAEP and PISA tests randomly select questions from a common universe of mathematics knowledge" (Hanushek et al., 2010, p. 10). The researchers argued that the high between-country correlation (.93) reported in a previous study between PISA mathematics scores and mathematics scores from the Trends in International Mathematics and Science Study (TIMSS) makes the two measures comparable. (Researchers had already linked TIMSS and NAEP mathematics in several studies—see Kilpatrick, 2011a, for more information.)

THE HAZARDS OF UNIDIMENSIONALITY

From a psychometric standpoint, rank ordering countries, states, or districts according to the percentage of test questions answered correctly is inappropriate when, as for NAEP, TIMSS, and PISA mathematics, the standard errors are greater than the differences in percentages (Stoneberg, 2005). Apparently, problems of scaling model misfit and differential item functioning also confound any ranking of countries on PISA scores (Kreiner & Christensen, 2014).

From a mathematics education standpoint, it makes no sense to treat NAEP and PISA as measuring the same construct. The NAEP eighth-grade mathematics test measures the mathematics proficiency deemed to be needed by U.S. students at that grade, whereas the designers did not intend for PISA mathematics to link to the school mathematics curriculum in any country. It measures the ability of 15-year-olds to apply the mathematics that they have learned to realistic situations. The study conflates "the results of two different tests that measure different domains of mathematics proficiency" (Kilpatrick, 2011a, p. 2).

> High between-country correlations do not at all imply that the same aspects of learning are being measured; instead, the measures are quite likely to be linked to similar learning assets (e.g., countries' wealth, prior educational level, and investment in schools). If high correlations were all it took to ensure comparability, then the PISA scores for mathematical literacy, scientific literacy, and reading literacy would all be comparable because their correlations at the country level in PISA are above .90. Would anyone want to claim that the PISA tests of scientific and reading literacy could therefore be used as measures of mathematical literacy? (Kilpatrick, 2011b, p. 1)

The Hanushek, Peterson, and Woessmann (2010) study is an extreme example of assuming that because a test is labeled *mathematics,* it must be essentially equivalent to any other test so labeled. Unfortunately, educational researchers working outside mathematics education all too commonly make such assumptions. They assume that because mathematics is a single domain, it is unidimensional.

As Orrill and Cohen (2016) point out in Chapter 7, the conceptualization of the domain to be assessed shapes the assessment. For example, proficiency in mathematics may look more or less unidimensional depending on one's purpose in measuring it. Orrill and Cohen observe that a certification assessment instrument such as the Praxis II for candidates who want to teach middle school mathematics simply provides a scaled score along a single dimension that allows each certification agency to set a cutoff. As long as the content of the assessment is representative of the mathematics taught in the middle grades, the dimensionality of the instrument is not much of an issue. Educators treat it as unidimensional for the purpose of selecting qualified candidates. In Chapter 5, Templin, Bradshaw, and Paek (2016) give a similar example of an end-of-course test for eighth-grade mathematics that would provide a single score "as a measure of a respondent's overall mathematical ability" (p. 100). The test would comprise items meant to be a representative sample of content from the course. The multiple

aspects of the course would, for the purpose of that assessment, be considered a single broad unidimensional construct. As Templin et al. observe, much depends on the conceptualization and operationalization of the constructs to be measured.

DIMENSIONALITY

In their discussion of extensions of item response theory (IRT) methods and models in Chapter 2, Bolt, Kim, Blanton, and Knuth (2016) note that "the concept of unidimensionality is statistical rather than substantive" (p. 34). In other words, one may simultaneously view proficiency in teaching, learning, or doing mathematics as having multiple components and nonetheless treat "it statistically as a unidimensional trait" (p. 34). Mathematics educators prefer to view proficiency in mathematics as a multidimensional phenomenon whether or not they treat it that way in a psychometric context.

In the 1960s, I participated in the National Longitudinal Study of Mathematical Abilities (NLSMA; see Begle & Wilson, 1970, and Howson, Keitel, & Kilpatrick, 1981, pp. 189–195, for details). The NLSMA had many weaknesses, but one thing it did right was not only to use *abilities* in the plural but also to borrow, construct, and administer a great many measures of mathematical proficiency. The director, E. G. Begle, and the NLSMA advisory board might have originally thought that they could determine how well students using different textbooks were learning mathematics by either using existing tests or by easily creating new ones. They quickly realized, however, that they would have to create and try out tests in the batteries used in the NLSMA in an extensive instrument-development effort. Those tests allowed comparisons on multiple dimensions among groups of students using different textbooks.

In the section of Chapter 8 that deals with "Analysis–Construct Discrepancies" (p. 163), Jacobson, Remillard, Hoover, and Aaron (2016) cite examples in which the dimensionality of the knowledge construct, as proposed by previous empirical and theoretical work, was not borne out when psychometricians applied standard psychometric techniques to assessment data. They observe that the process of developing and testing the assessment instruments led the researchers to revise how they had conceptualized the domain. One could just as plausibly argue, however, that the assessment instruments—and not just the conceptualizations—also needed revision. In fact, in the two examples that Jacobson et al. give, one can see the revision process going both ways, with the reformulation of constructs and the writing of new assessment items.

In Chapter 1, Izsák and Templin (2016) point out that from a psychometric point of view, unidimensional models of mathematical topics are "much more common" (p. 8) than multidimensional models. I would argue that every topic in school mathematics is, as they indicate some significant topics are, "intricate and multifaceted" (p. 11) and therefore ought to have a multidimensional model if possible. One problem seems to be not that such models do not exist but rather that if the researchers model the dimensions as continuous

variables, reliable measurement of each dimension would demand an impractical number of assessment items. Izsák and Templin also point out that using a unidimensional model of growth in understanding a specific mathematical topic might be misleading. They suggest that models based on categorical latent variables and models based on a combination of continuous and categorical variables might help researchers in mathematics education move beyond the unidimensionality issue.

In Chapter 6, Kersting, Stevenson, and Chen (2016) consider dimensionality in the context of instrument design and development. They deal with the unidimensionality issue, in part, by introducing the notion of *essential unidimensionality*—given that "real assessment data are almost never truly unidimensional" (p. 121). Assessment data are essentially unidimensional if a bifactor analysis reveals one predominant factor in the data together with weak orthogonal group factors. Kersting et al. observe that multidimensionality can arise among assessment items (when different items measure different latent traits) or within items (when individual items measure multiple latent traits). They point out that "dimensionality is largely a design choice" (p. 124), which I would interpret as applying both to the conceptualization of the constructs and to the creation of the assessment instruments.

Interpreting the statistical analysis of their data, Kersting et al. (2016) argue that "the data have both multidimensional and unidimensional characteristics" (p. 134) and can be modeled either way. This argument suggests that even if mathematical phenomena are multidimensional, the dimensionality of the constructs used to model them and the instruments used to assess those constructs may be to some degree arbitrary—dependent not just on existing theory and research but also on how researchers have conceptualized the constructs, developed the assessment instruments, and employed statistical procedures such as factor analysis.

TRADE-OFFS

A theme that runs through the preceding chapters in this monograph, just as it did during the conference from which the monograph arose, concerns the trade-offs that arise in the use of IRT models, diagnostic classification models (DCMs), and their combinations. In Chapter 4, Tatsuoka et al. (2016) note that trade-offs exist between "the desired level of detail about examinees' reasoning and the extent of the covered mathematical terrain" (p. 75). Given practical limitations on the size of the sample of participants and on the length of assessment instruments, Izsák and Templin (2016) identify two trade-offs in Chapter 1: (a) between the ability to order one's data with precision (IRT models) and the ability to assess multiple dimensions simultaneously (DCMs) and (b) between model complexity and sample size. In Chapter 3, de la Torre, Carmona, Kieftenbeld, Tjoe, and Lima (2016) see these trade-offs as "a tension between depth and breadth" (p. 67) regarding what is possible in a measurement situation.

They also point out that although DCMs provide researchers with "a rich set of different grain-sized descriptors of domain-specific knowledge" (p. 55), the time demanded to construct and improve the valid Q-matrices and assessment instruments can be considerable—an observation that Tatsuoka et al. (2016) made in Chapter 4.

Much of the trade-off issue appears to involve time: the time needed to develop and refine theories, constructs, and assessment instruments; the time that might be necessary to generalize theories and constructs to other contexts and to address issues of equity and social aspects of mathematics learning; the limited time that students and teachers have available for taking assessments; and the added time that might be necessary to score items that require constructed responses. In Chapter 4, Tatsuoka et al. (2016) point out that adaptive testing can achieve considerable savings in the time needed for assessment administration, and it is also true that scoring methods continue to become more efficient. However, the time needed for researchers in mathematics education and psychometricians to learn to work together and to conduct the research studies needed to support their work is not likely to diminish.

CONCLUSION

Researchers in mathematics education have a special perspective on measurement. From one angle, measurement is the process that they use to determine the nature and extent of the teaching, learning, and doing of mathematics in which their research participants engage. Much of that process involves drawing inferences about mental objects or activities, and issues of reliability and validity loom large whether the measurement involves scaling a quantity or sorting a quality. From another angle, for mathematics educators, measurement is a curriculum strand that begins with preschoolers comparing the lengths, areas, volumes, times, or other attributes of real objects and events that they will eventually learn to measure. The measurement strand continues through the grades, extending as far as graduate students generalizing their intuitive notions of length, area, and volume when they learn measure theory. In the school curriculum, measurements are more likely to be of physical objects than mental objects or activities; and issues regarding such concepts as units, iteration, tiling, proportionality, and additivity loom much larger than reliability or validity.

Consequently, researchers in mathematics education deal with, on the one hand, the measurement of teaching, learning, and doing mathematics and, on the other hand, the teaching, learning, and doing of measurement. In either case, they need to understand that all measurement relies on a mathematical model of the constructs being measured. Researchers in mathematics education are well positioned to understand the assumptions on which those models depend, but they may also be especially prone to question the assumptions. Recent developments in psychometrics have the potential to help those researchers clarify the measurement models.

Interaction between theory and practice is an issue both in mathematics education and in psychometrics. Just as mathematics educators can help psychometricians avoid treating all mathematics tests as equivalent, psychometricians can help mathematics educators avoid drawing unwarranted conclusions about the results of those tests. The present monograph should offer the reader hope that, by working together however long it may take, scholars from the two fields can improve their theories as well as their practices.

REFERENCES

Begle, E. G., & Wilson, J. W. (1970). Evaluation of mathematics programs. In E. G. Begle (Ed.), *Mathematics education* (69th Yearbook of the National Society for the Study of Education, Part 1, pp. 367–404). Chicago, IL: University of Chicago Press.

Bolt, D. M., Kim, J.-S., Blanton, M., & Knuth, E. (2016). Applications of item response theory in mathematics education research. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 31–52). *Journal for Research in Mathematics Education* Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

de la Torre, J., Carmona, G., Kieftenbeld, V., Tjoe, H., & Lima, C. (2016). Diagnostic classification models and mathematics education research: Opportunities and challenges. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 53–71). *Journal for Research in Mathematics Education* Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2010, November). *U.S. math performance in global perspective: How well does each state do at producing high-achieving students?* (PEPG Report No. 10–19). Cambridge, MA: Harvard's Program on Education Policy and Governance and *Education Next*. Retrieved from http://www.hks.harvard.edu/pepg/PDF/Papers/PEPG10-19_HanushekPetersonWoessmann.pdf

Howson, G., Keitel, C., & Kilpatrick, J. (1981). *Curriculum development in mathematics*. Cambridge, United Kingdom: Cambridge University Press. doi:10.1017/CBO9780511569722

Izsák, A., & Templin, J. (2016). Coordinating conceptualizations of mathematical knowledge with psychometric models: Opportunities and challenges. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 5–30). *Journal for Research in Mathematics Education* Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Jacobson, E., Remillard, J. T., Hoover, M., & Aaron, W. (2016). The interaction between measure design and construct development: Building validity arguments. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 155–173). *Journal for Research in Mathematics Education* Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Kersting, N. B., Stevenson, P. A., & Chen, M.-K. (2016). Examining and understanding dimensionality in the context of instrument development: Considerations from the Classroom Video Analysis instrument measuring usable teaching knowledge in mathematics. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 119–138). *Journal for Research in Mathematics Education* Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Kilpatrick, J. (2011a). Review of math performance in global perspective [Review of the report *U.S. math performance in global perspective: How well does each state do at producing high-achieving students?* by E. A. Hanushek, P. E. Peterson, & L. Woessmann]. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/files/TTR-Global-math_0.pdf

Kilpatrick, J. (2011b). *Responding by not responding: A reply to Paul E. Peterson*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/files/Response to Peterson.pdf

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231. doi:10.1007/s11336-013-9347-z

Orrill, C. H., & Cohen, A. (2016). Purpose and conceptualization: Examining assessment development questions through analysis of measures of teacher knowledge. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 139–153). *Journal for Research in Mathematics Education* Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Stoneberg, B. D. (2005). Please don't use NAEP scores to rank order the 50 states. *Practical Assessment, Research & Evaluation, 10*(9). Retrieved from http://pareonline.net/getvn.asp?v=10&n=9

Tatsuoka, C., Clements, D. H., Sarama, J., Izsák, A., Orrill, C. H., de la Torre, J., … Tatsuoka, K. K. (2016). Developing workable attributes for psychometric models based on the Q-matrix. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 73–96). *Journal for Research in Mathematics Education* Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.

Templin, J., Bradshaw, L., & Paek, P. (2016). A comprehensive framework for integrating innovative psychometric methodology into educational research. In A. Izsák, J. T. Remillard, & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations* (pp. 97–117). *Journal for Research in Mathematics Education* Monograph Series No. 15. Reston, VA: National Council of Teachers of Mathematics.