The 2015 Annual Meeting of AERA
April 16-20, 2015
Chicago, IL

Are we Measuring it Accurately: Mathematical Literacy in the PISA Context?

Adem Ekmekci
Rice University
ekmekci@rice.edu

Large-scale assessments' validity and reliability matters more than we think because accurate interpretations and sound judgments about students' mathematical literacy depend on them. This study investigates the test dimensionality of PISA's mathematics assessment to see the alignment between the cognitive framework for mathematical literacy and actual results (students' responses to mathematics items) as one important type of evidence for validity. Our results show that the multidimensional structure of mathematical literacy is not reflected well-enough in the mathematics items. However, longitudinal consistency of the results across cycles provides evidence toward PISA's reliability. These results have important implications for the way mathematical literacy is assessed from mathematics education and psychometric perspectives.

Introduction

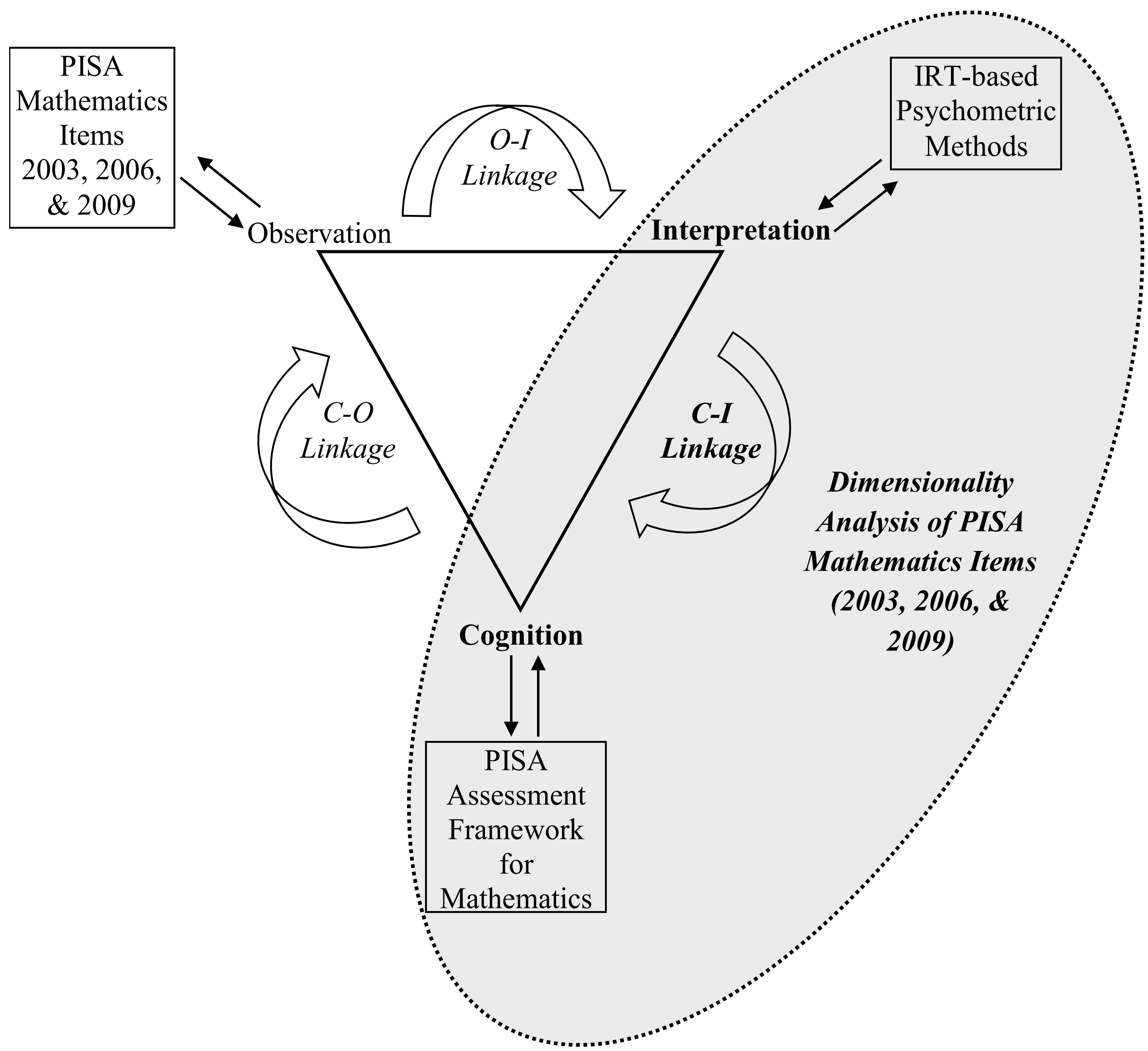
This research relates to validation of assessments in mathematics at a large-scale. The context for this study is the Programme for International Student Assessment (PISA). The goal is to study the conformity between the cognitive framework provided for mathematical literacy (the intended structure,) and the statistical structure of students' responses to individual items in PISA over the years. National Research Council (NRC) recommends the three components of an assessment design: cognition, observation, and interpretation, need to be coordinated in a consistent and integrated way (not isolated from each other) (Pellegrino, Chudowsky, & Glaser, 2001). This study focuses on the alignment between the theoretical framework for cognition and the score interpretation framework provided for PISA mathematics assessment.

There are a limited number of studies investigating the connection between the assessment framework and results. Schwab (2007) found that the multidimensional nature of PISA's science framework was not reflected well enough in the items. Author (2012) studied the students' responses to PISA 2003 mathematics items and detected unidimensionality for the U.S. student population. Similar results were found in Author's (2014) study analyzing mathematics items with 2009 students. However, this study extends prior work and combines them in order to understand better the complexities of assessing mathematical literacy at a large scale. This study presents a dimensionality analysis of PISA 2003, 2006, and 2009 (all implementation cycles prior to the second wave with mathematics being the major domain again in 2012, whose results have recently been released) mathematics items using **all** students' responses to individual items from **all** Organisation for Economic Co-operation and Development (OECD) countries.

Theoretical Framework

NRC outlines an assessment design framework in *Knowing What Students Know*. This framework proposes the integration of three components in assessment design that can be represented by a triangle, with each corner representing *cognition*, *observation*, and *interpretation* (Pellegrino et al., 2001). Cognition refers to the model of student learning in the domain, or mathematical literacy for our study; observation consists of the evidence provided by the student of the assessed construct; and interpretation entails making sense of this evidence. This *triangle* representation signifies the idea of a need for interconnectedness, consistency, and integrated development of the three elements, as opposed to having them as isolated from each other. Based on the recommendations for research outlined in the NRC's assessment report (e.g., conformity between assessment framework-*cognition*, and assessment results-*interpretation*), this study aims to investigate dimensionality of PISA mathematics items (Figure 1).

Figure 1. Conceptual framework



The concept of mathematical literacy gained crucial importance especially in the 80's. Since then, the standards that had been once considered for literacy (being able to read and write) also began to be considered for mathematical literacy (Jablonka, 2003; Moses & Cobb, 2001). What motivates this study is the perception and reflection of mathematical literacy in the assessment context, in particular large-scale assessments whose results may have serious impact on education systems. Mathematical literacy is defined and viewed in the literature as a multidimensional construct consisting of distinguishable but related components rather than single, general mathematics ability. Some math educators focus on proficiencies or competencies (e.g., Kilpatrick, Swafford, & Findell, 2001) when defining mathematical literacy, while others describe knowledge and skills (e.g., Ojose, 2011). Others situate mathematical literacy according to its connection to real life situations (i.e., context) (e.g., Steen, 2001). So, there appears to be more than one dimension and more than one approach in composing mathematical literacy as discussed in the mathematics education field.

The PISA views mathematical literacy as a multidimensional construct composing of *content, processes,* and *context*, each having three to four sub-dimensions (OECD, 2003). The goal of this study is to show how and to what extent this complex multidimensional nature of assessment framework is reflected on the actual tests by investigating the dimensional structure of the PISA mathematics items using the student responses, and to monitor this correspondence over three implementation cycles (2003, 2006, and 2009).

One of the most powerful ways to explore the connection and conformity between the framework for mathematical literacy and its assessment is dimensionality analysis and requires an understanding of dimensionality concept. Test dimensionality could be informally defined as "*the minimum number of examinee abilities measured by the test items*" (Tate, 2002, p.182). If assessment items form a unidimensional structure, then this set of items is said to be measuring one attribute of a construct. If an assessment is said to be measuring several important aspects of a construct, then it is supposed to have a multidimensional structure. Dimensionality relates to central issues in development and use of large-scale assessments such as content validity, construct validity, score reliability, and test fairness. For example, unidimensionality is the basic assumption of measurement models (Hattie, 1985) and is required for construct validity (Rubio, Berg-Weger, & Tebb, 2001). However, it is sometimes the case that a test that is intended to be unidimensional may unintentionally be measuring more than one latent variable. Therefore, analyzing the dimensionality of an assessment it important and required to ensure accurate interpretations of its results.

Methods

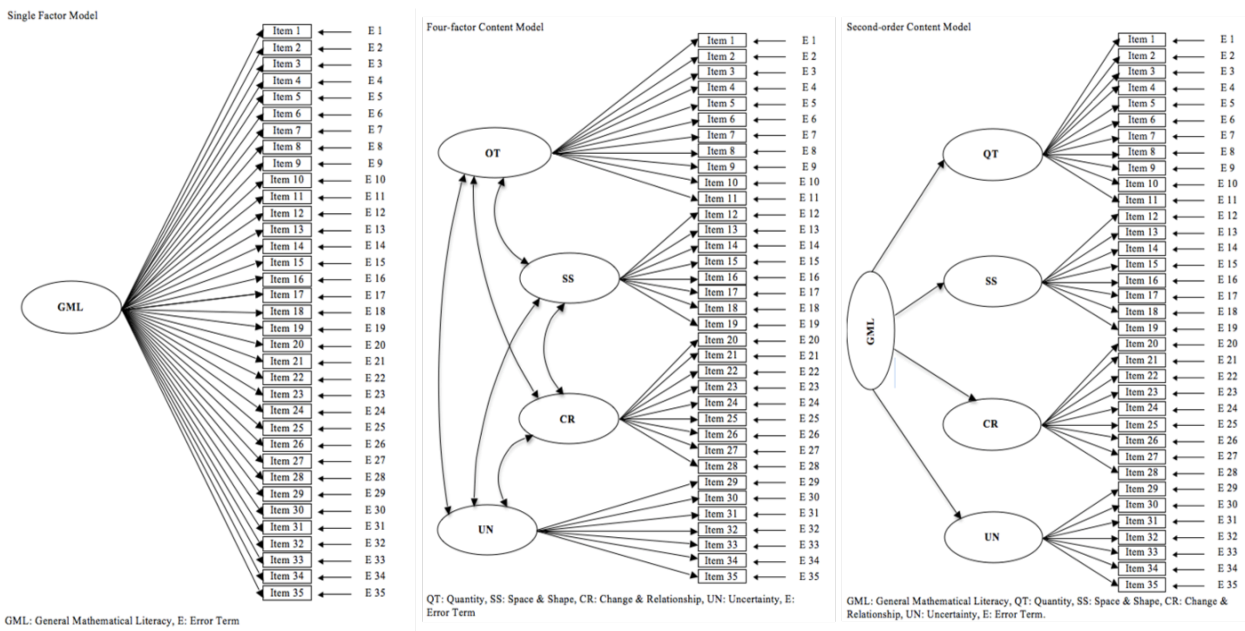
The two research questions guided this study are:

1. What is the best representation for the dimensional structure of the PISA mathematics items for implementation cycles 2003, 2006, and 2009?
2. How does the dimensional structure of the PISA mathematics items change over time?

This study entails a secondary analysis of the dataset from the OECD's PISA database. The data includes student responses to individual mathematics items from 32 OECD countries in PISA 2003, 2006, and 2009 cycles. Confirmatory factor analysis (CFA) techniques are conducted to investigate the dimensionality of PISA mathematics items. Seven CFA models were developed based on the mathematical literacy dimensions described in OECD's assessment framework for mathematical literacy. These models include one unidimensional model, three (content, process, and context) correlated factor (1-level) models, and three (content, process, and context) higher order factor (2-level) models. Figure 2 provides models for content dimension as an example. Each CFA model was tested with the student responses to mathematics items from 32 OECD countries for each implementation cycle.

For each three implementation cycles, the first model (single-factor) hypothesizes that PISA mathematics items measure a single construct labeled as general mathematical literacy (GML). The second type of model, correlated-factors model or 1-level model, (Models 2-4) hypothesizes that the PISA mathematics items helps explain mathematics knowledge, competencies, and skills in terms of correlated factors of related dimension (content, process, or context) as the latent constructs. The third type of model, second-order model or 2-level model, (Models 5-7) hypothesizes that the PISA mathematics items measure GML (level-2 factor) by factors (the level-1 latent variables) of related dimension (content, process, or context).

Figure 2. CFA models: One single factor model (left), three correlated-factors models (middle), three second-order models (right)



The first research question explores the models that best represent the dimensionality of response data in different years. Different structural models were built based on the PISA mathematical literacy framework. In order to explore the extent to which each model would fit the student responses to PISA mathematics items and to find the best fitting model, some statistical indices were needed. Goodness of fit indices (GFIs) obtained from CFA analyses such as comparative fit index (CFI), the Tucker-Lewis index (TLI), and root mean square error of approximation (RMSEA) were used to evaluate model-fit. Moreover, DIFFTEST (alternative version of chi-square difference testing modified for WLSMV estimator) and ΔGFI methods were used to compare models within each three main dimensions (content, process, and context) for each implementation cycle.

The second research question, as an extension of the first one, investigates whether the PISA mathematics assessment has stability in terms of dimensional structure. It is expected that that model comparison results would be stable across different cycles. Therefore, models were evaluated across three implementation cycles, addressing the longitudinal aspect of dimensionality of PISA mathematics items.

Results

The statistical software Mplus 6.12 (Muthen & Muthen, 1998-2011) was used to conduct confirmatory analyses (with WLSMV estimator for categorical data). The resulting estimated values for the 2009 implementation cycle are given in Table 1 below. The estimates for 2003 and 2006 cycles are very similar to 2009 cycle whose summary of the results are provided below. All seven models, including the unidimensional and multidimensional, for the PISA mathematics items were found to be a good fit for all three implementation cycles: 2003, 2006, and 2009. In other words, the responses to the mathematics items do not contradict any of the models proposed for the dimensionality of PISA mathematics framework. However, high correlations between latent factors in level-1 models and high latent factor loadings in level-2 models further support the unidimensionality (Figure 3).

Model comparison results are very consistent for the content, process and context dimensions across the cycles. Across different dimensions, there are slight but in significant differences. This supports the reliability of PISA mathematics assessment. DIFFTEST and ΔGFI analyses revealed that the unidimensional model performed better with the responses dataset for PISA mathematics items (Figure 4).

Table 1. Model fit indices for 2009 cycle

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
χ^2 value	743.5	711.2	741.6	729.4	713.7	742.6	731.9
d.f.	560	554	557	554	556	559	556
p-*value*	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.000
CFI/TLI							
CFI	0.980	0.983	0.980	0.981	0.983	0.980	0.981
TLI	0.979	0.982	0.979	0.980	0.982	0.979	0.980
RMSEA							
Estimate	0.005	0.004	0.005	0.005	0.004	0.005	0.005
90% C.I.	[0.004, 0.005]	[0.003, 0.005]	[0.004, 0.006]	[0.004, 0.005]	[0.003, 0.005]	[0.004, 0.005]	[0.004, 0.005]
Prob. ≤ 0.05	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Conclusions

Findings from this study demonstrate that the connection between the assessment framework and the statistical structure of mathematics items is rather a weak one. That is, the intended multidimensional nature of mathematics items is not reflected well enough in the student responses. This same conclusion holds longitudinally over the different implementation cycles. This study used the most robust tools identified in the literature for analyzing test dimensionality. Moreover, PISA is one of the most widely recognized and respected assessments in the world,

Figure 3. Correlations between factors in each dimension for 2009 implementation cycle

Content

	QT	SS	CR	UN
QT	1			
SS	0.876	1		
CR	0.917	0.898	1	
UN	0.883	0.860	0.962	1

Process

	REP	CON	REF
REP	1		
CON	0.988	1	
REF	0.986	0.981	1

Context

	PER	EDOP	PUB	SCI
PER	1			
EDOP	0.948	1		
PUB	0.937	0.907	1	
SCI	0.947	0.974	0.941	1

Figure 4. Model comparison results

2003

Content: 1F-GML Model > 2-Level Model > 1-Level Model

Process: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model

Context: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model

2006

Content: 1F-GML Model ≥ 1-Level Model ≥ 2-Level Model

Process: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model

Context: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model

2009

Content: 1F-GML Model > 2-Level Model > 1-Level Model

Process: 1F-GML Model ≥ 2-Level Model ≥ 1-Level Model

Context: 1F-GML Model > 2-Level Model > 1-Level Model

having a well-articulated and comprehensive mathematical literacy framework and a robust psychometric design. Yet, the major components of its assessment design are not at an expected level of corroboration. This has important implications for mathematics education, measurement, and psychometrics fields. The next step in this line of work is to test the dimensionality of PISA 2012 that has been recently released. Given that the mathematics was the major domain once again with improved items, it would be interesting to see whether the multidimensionality of the framework was reflected good enough in the 2012 mathematics items.

Significance of the Study

The dimensionality of PISA's mathematical literacy assessment over the course of first mathematics wave (2003, 2006, and 2009) has not been undertaken before. Thus, this investigation is an important contribution to the study of its validity. Moreover, assessing the dimensionality of PISA mathematics items is needed to understand the relationship between the important components (assessment triangle) of PISA assessment design for mathematical literacy. Prior studies have set the ground but have left a gap in assessing dimensionality of PISA mathematics assessment. This study has the potential to fill in this gap. The significance of this study comes from the need to provide evidence for validation process of PISA mathematical literacy assessment.

The author argues that psychometric methods currently being used for large-scale assessments (e.g. Rasch models) might be too limiting to provide evidence for the types of constructs the field of mathematics education is interested in and in need of assessing. The context in this study is mathematics, but it may apply to other fields such as reading and science as well. An important implication for the field of mathematics education is that the field is in high need of new assessment designs that would bring in other views on mathematics literacy - beyond those addressed in PISA, together with more current psychometric models that allow for assessment of multidimensional constructs, and therefore providing a more encompassing perspective and more valid assessments, especially those that are implemented at a large-scale and that have such high stakes decisions based on these results.

References

Ekmekci, A., & Carmona, G. (2012). Mathematical literacy assessment design: A multivariate analysis of PISA 2003 mathematics items in the U.S. In Van Zoest, L. R., Lo, J.-J., & Kratky, J. L. (Eds.), *Proceedings of the 34th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education,* (p. 390). Kalamazoo, MI: Western Michigan University.

Ekmekci, A., & Carmona, G. (2014). Studying mathematical literacy through the lens of PISA's assessment framework. In Nicol, C., Liljedahl, P., Oesterle, S. & Allan, D. (Eds.), *Proceedings of the 38th Conference of the International Group for the Psychology of Mathematics Education and the 36th Conference of the North American Chapter of the Psychology of Mathematics Education - Vol. 2* (pp. 441-448). Vancouver, Canada: PME.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Jablonka, E. (2003). Mathematical literacy. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Second International Handbook of Mathematics Education* (pp. 77-104). Berlin: Springer.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC.: National Academy Press.

Moses, R. P., & Cobb, C. E. (2001). *Radical equations: Civil rights from Mississippi to the algebra project*. Boston, MA: Beacon.

Muthén, L.K. and Muthén, B.O. (1998-2011). *Mplus, Version 6.12 (Computer Software).* Los Angeles, CA: Muthén & Muthén.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds). (2001). (2001). *Knowing what students know: The science and design of educational assessment*. National Research Council, Washington, DC: The National Academy Press.

Ojose, B. (2011). Mathematics literacy: Are we able to put the mathematics we learn into everyday use? *Journal of Mathematics Education, 4*(1), 89-100.

Organisation for Economic Co-operation and Development (OECD). (2003). *The PISA 2003 assessment framework - mathematics, reading, science, and problem solving knowledge and skills*. Paris: OECD Publishing.

Rubio, D. M., Berg-Weger, M., & Tebb, S. S. (2001). Using Structural Equation Modeling to Test for Multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(4), 613-626.

Steen, L. A. (2001). *Mathematics and democracy: The case for quantitative literacy*. Princeton, NJ: National Council on Education and the Disciplines.

Schwab, C. J. (2007). *What can we learn from PISA? Investigating PISA's approach to scientific literacy*. (Unpublished Dissertation). University of California, Berkeley, U.S.

Tate, R. (2002). Test Dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation*. Mahnah, New Jersey, London: L. Erlbaum.