Collaboration Between Content Experts and Assessment Specialists: Using a Validity Argument Framework to Develop a College Mathematics Assessment

Amanda Brijmohan
University of Toronto

Gulam A. Khan
University of Toronto

Graham Orpwood
York University

Emily Sandford Brown
Sheridan College

Ruth A. Childs
University of Toronto

Canadian Journal of Education / Revue canadienne de l'éducation 41:2 (2018)

www.cje-rce.ca

Abstract

Developing a new assessment requires the expertise of both content experts and assessment specialists. Using the example of an assessment developed for Ontario's Colleges Mathematics Assessment Program (CMAP), this article (1) describes the decisions that must be made in developing a new assessment, (2) explores the complementary contributions of content experts and assessment specialists, and (3) illustrates how the use of a validity argument framework can support collaboration in assessment development. The authors conclude that the validity argument framework facilitated effective collaboration between content experts and assessment specialists, and suggest that this approach may help other collaborators pursue transparent and effective assessment development.

Keywords: assessment, test development, mathematics assessment, colleges

Résumé

Le développement de nouveaux outils d'évaluation nécessite l'expertise tant d'experts de contenu que de spécialistes en évaluation. En utilisant l'exemple d'un outil d'évaluation développé pour l'Ontario's Colleges Mathematics Assessment Program (CMAP), cet article (1) décrit les décisions qui doivent être prises lors de l'élaboration de nouveaux outils d'évaluation, (2) explore les apports complémentaires des experts de contenu et des spécialistes en évaluation, et (3) illustre comment l'utilisation d'un modèle d'argumentation de la validité peut favoriser le développement d'évaluations. Les auteurs concluent qu'un modèle d'argumentation de la validité simplifie le processus de collaboration efficace entre les experts de contenu et les spécialistes en évaluation, et suggèrent que cette approche puisse aider d'autres collaborateurs à poursuivre l'élaboration d'évaluations authentiques et significatives.

Mots-clés : outils d'évaluation, développement d'évaluation, évaluation en mathématique, collèges

Introduction

For more than a decade, poor mathematics performance of Ontario college students has been an area of concern. The College Student Achievement Project (CSAP) found, for example, that one-third of college students were "at-risk" of not completing the requirements for their program due to poor mathematics performance in their first semester (CSAP Team, 2015). This is especially problematic as finding success in a variety of job fields requires numeracy skills (Gal & Tout, 2014; Hoyles, Wolf, Molyneux-Hodgson, & Kent, 2002; Organisation for Economic Co-operation and Development [OECD], 2013; Steen, 2001). Therefore, understanding the gaps in mathematics knowledge and skills of students entering college programs is important if colleges are to prepare those students for success in college programs and in future employment (Cohen & Brawer, 2003; Ngo & Melguizo, 2016).

To identify gaps in students' knowledge and skills, some colleges administer written assessments to incoming students. Developing such assessments requires input from content experts, as well as assessment specialists. However, how their contributions can be combined to create a quality assessment is not always obvious. Using the example of an assessment developed for Ontario's Colleges Mathematics Assessment Program (CMAP), this article (1) outlines the decisions that must be made in developing a new assessment, (2) explores the complementary contributions of content experts and assessment specialists, and (3) illustrates the use of a validity argument framework for collaboration in assessment development. Although the example is of a mathematics assessment, the approach described is not specific to mathematics.

Addressing Gaps in Students' Mathematics Knowledge and Skills

When gaps in students' mathematics knowledge and skills are identified, some programs will direct students to a remedial course (i.e., a course that is not required in a program, but is intended to prepare students for the courses that are required). In Ontario, such courses have been well subscribed: In 2012, for example, 25.8% of students enrolled in a college-level mathematics course were enrolled in a remedial course, an increase from

17.3% in 2008 (CSAP Team, 2015). The assumption is that placing students into appropriate remedial courses will provide enough support for those students to subsequently succeed in their regular courses. Though research suggests that remediation can make up for mathematics preparedness gaps for some students (Bahr, 2008), not all students have been shown to benefit (Bahr, 2013; Bailey, 2009). Bahr (2013) highlights poor retention in remedial courses as a barrier to closing mathematics skills gaps, noting that not all students who are enrolled in remedial courses actually finish. Bailey (2009) also echoes this finding and, further, finds that of those who do complete their remediation courses, many do not proceed to enroll in the subsequent college-level mathematics courses. Bailey, Jeong, and Cho (2010) found in a review of the research literature that the same patterns have been observed in other college remediation contexts as well.

Identifying Gaps with Placement Tests

At many colleges, recommendations for remedial courses rely on the results of placement tests developed by the college's faculty and staff. These tests are intended to determine whether students have sufficient mathematical knowledge and skills to succeed in the mathematics courses in their program or need to take a remedial course; if remediation is needed, such tests may reveal what level of remediation is appropriate (Ngo & Melguizo, 2016). How well the tests inform placement decisions has been questioned, however. Fields and Parsad (2012) found that US community colleges had varied placement and assessment policies and used different cut-off scores for the same tests. Furthermore, some research suggests that mathematics placement tests may be placing close to 25% of students into inappropriate remedial mathematics courses (Scott-Clayton, Crosta, & Belfield, 2014).

In addition to being used to inform placement decisions, some tests are also used to decide whether a student will be offered admission to a college program. For programs that receive more applications than available spaces, weak performance on a mathematics placement test may be a reason for not offering admission to an applicant.

Developing Assessments

Given the stakes of these decisions for students, the quality of the placement tests is important. Creating a placement test requires a series of decisions, including what domains of knowledge should be assessed, how they should be assessed, and what the standards of mathematics performance should be for students beginning each program (Schmeiser & Welch, 2006). Making such decisions, however, is anything but straightforward. For example, Melguizo, Kosiewicz, Prather, and Bos (2014), based on their examination of the local assessment and placement policies of one community college, concluded that "faculty and administrators possess little knowledge about which test works most effectively to place students, how to rigorously evaluate cut scores, and which multiple measures can adequately address short comings inherent in current placement tests" (pp. 714–716). Melguizo and colleagues note that what should be most concerning to researchers, policymakers, and institutions is that students who have similar levels of mathematics knowledge and skills but attend colleges with different placement tests or criteria could potentially be placed into different levels of remedial courses or denied admission to one program, while gaining admission to a comparable program at another college.

Developing a mathematics assessment to inform placement and admission decisions requires the content expertise of mathematics educators. Content expertise consists of both content knowledge (understanding what knowledge and skills students will need in their programs) and pedagogical content knowledge (understanding how students learn, common difficulties they experience, and the relationship between their skills and knowledge) (Ball et al., 2007; Shulman, 1986). Given the importance of the decisions that will be based on the assessment results, it is also essential to involve assessment specialists with advanced training in measurement, assessment development and validation. Effective collaboration between these two groups of experts is necessary if the resulting assessment is to be of high quality. Reaching a shared understanding of the implications of decisions about an assessment's content, format, scoring, and reporting can be difficult, however.

Mislevy, Steinberg, and Almond (2003) have offered one possible framework for discussing the implications of decisions in assessment development. In evidence-centered design (ECD), those developing an assessment will work together to define a *student*

model specifying the constructs to be measured, a *task model* with schemas for collecting evidence about the constructs, an *evidence model* specifying how the tasks will be scored and those scores combined, an *assembly model* specifying how tasks will be selected for a test form, and a *presentation model* dealing with details of task presentation and test administration. Together, these models form a *conceptual assessment framework*.

Mislevy and colleagues (2003) emphasize that one of the goals of ECD is to make explicit "the connections among an assessment's purpose, a conception of proficiency in the domain, the evidentiary argument, the design of the assessment elements, and operational processes" (p. 2). Indeed, in a discussion of their experiences using ECD, Hendrickson, Ewing, Kaliski, and Huff (2013) identified two primary benefits: "(1) assessments better reflect and measure what is taught and valued in the classroom, and (2) resulting score inferences are strongly supported by an evidentiary argument" (p. 4). However, they also pointed out the complexity and resource requirements of using ECD. To address the complexity, they offered a checklist for use in confirming that each part of the framework has been addressed; with respect to the resource requirements, they believe that "the use of ECD will become less resource intensive once it is employed more broadly in assessment design and development, and we collectively identify pathways through the current challenges" (p. 5).

While the ECD has the potential to help assessment development teams think about the implications of their decisions, the complexity and resource demands noted by Hendrickson and colleagues (2013) are likely to present difficulties for teams that do not regularly develop assessments together. Indeed, an alternative approach is hinted at in their observation that "justifying the resources required for ECD falls largely on the argument that the benefits include an improved validity argument" (p. 24). A few test developers have, instead of using ECD as a framework for decision making, worked to make the implications of test development decisions clear by linking the decisions to parts of a validity argument.

Validity Arguments

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) defines validity as "the degree to which evidence and theory support the

interpretations of test scores for proposed uses of tests" (p. 11). The specifics of what is meant by validity and what constitutes validity evidence have been debated for many decades (see Newton and Shaw, 2014, for a comprehensive history); however, in the past 15 years, the practice of organizing evidence into a validity argument has become increasingly common. Developing a validity argument is typically seen as involving two steps, summarized by Kane (2016) as "(a) specify the claims that are to be based on test scores, as an interpretation/use argument, or IUA, and (b) evaluate the plausibility of these claims using appropriate methods and evidence in a *validity argument* [emphasis in the original]" (p. 309). The IUA usually includes at least three inferences or claims: (a) a scoring inference, which concerns the relationship of the individual items to the test score or scores, (b) a generalization inference, concerning whether the score(s) are specific to these items or could have been obtained from other samples of items from the domain, and (c) an extrapolation inference, concerning the relationship of the score(s) to what the developers intended the test to measure (Kane, 2001, 2013).

Though Kane's IUA is widely used in validity research, two notable critiques of the framework come from work done by Schilling and Hill (2007), best known for their use of validity arguments in mathematics education research. While Kane's IUA framework does not differentiate between assumptions and inferences, Schilling and Hill (2007) argue that test assumptions and inferences are distinct entities that should be made explicit in the validity argument: Assumptions are driven by an underlying theory, while inferences are empirically testable consequences of the test. Second, Kane's IUA is considered to be hierarchically structured, where the scoring, generalization, and extrapolation steps are sequential. Schilling and Hill (2007) propose a variation of this approach, grouping the inferences and assumptions into three components: (1) elemental, (2) structural, and (3) ecological. At the elemental level, validity evidence for assumptions and inferences is gathered to check for item consistency. At the structural level, the domain structure and the subscales of the test are considered in support of proposed assumptions and inferences. At the ecological level, the external structure of the test is considered by looking at how test scores or subscores relate to external variables.

In their application of this validity framework to mathematics education research, Schilling and Hill (2007) examined the Mathematical Knowledge for Teaching (MKT) scales. Building on Kane's IUA, Schilling and Hill (2007) suggested ways to systematically gather validity evidence at each level. For example, at the elemental level, evidence

gathered to check item consistency may be through item discrimination analyses, or through think-aloud interviews to examine whether item responses are consistent with the test-takers' reasoning. At the structural level, evidence may be gathered to examine subscale dimensionality through factor analysis, or through cognitive interviews to understand shared reasoning among test takers, and whether this is consistent with the theory underpinning the constructs being assessed. At the ecological level, test scores or subscores may be correlated with external measures that theoretically measure the same constructs.

Developing a Framework for Assessment Development Decisions

In developing the CMAP, we built on Schilling and Hill's (2007) work, using their three categories to relate inferences and evidence to decisions made in test development. We extended their work by developing a presentation that emphasizes the role of the test development decisions in the validity argument and expresses the validity argument in nontechnical language. The goal of the framework was to facilitate collaborative decision-making among members of a team that included both mathematics educators and assessment specialists. The mathematics educators had content expertise, encompassing both content knowledge and pedagogical content knowledge, and experience teaching mathematics and developing mathematics curriculum at the secondary and college levels. The assessment specialists had expertise in psychometrics, and experience in articulating how each piece of evidence collected contributed to the validation process.

The Colleges Mathematics Assessment Program (CMAP) is a basic numeracy test developed by a group of Ontario colleges to test whether new students required remedial mathematics courses before beginning their intended program of study (Orpwood & Brown, 2014). In addition to its use as a placement test, CMAP is also intended, in a longer version, to help prospective students evaluate their own strengths and weaknesses in mathematics.

Developing the CMAP required decisions about the content and format of the test, and how scores would be computed and reported. Without a framework for considering the likely effects of those decisions, and for planning how these effects might

be investigated—that is, what evidence might be gathered during test development and later—it can be difficult to know which decisions are appropriate for the particular assessment. Tables 1 and 2 provide such a framework.

Table 1 shows the relationship between decisions the team had to make in developing the assessment and inferences in a validity argument. Table 2 provides examples of evidence to support each inference. In some cases, the realization that it would not be possible to obtain appropriate evidence shaped a decision. For example, some of the items contributed by the colleges for inclusion in the item bank had been designed to measure multiple content areas simultaneously. When we considered what evidence would be needed to support the inclusion of such items, however, we realized that neither the placement nor diagnostic forms of the assessment would have sufficient numbers of items or responses from students for the complex analysis models needed to combine responses across items that measure multiple content areas. Therefore, these items were reviewed and, where possible, edited to align with a single content area. This is an example of how the decisions determined the inferences and the required evidence in the validity argument.

Table 1. Test development decisions and contributions from team members

In response to the question…	the team decided…	based on…
What knowledge and/or skills should be tested?	Content: Whole numbers, arithmetic, integers, decimals, fractions, ratio and proportion, percents, basic algebra, measurement. Performance expectations: Knowing, applying, reasoning	The content experts' review of the provincial high school curriculum in relation to the mathematics required in college Business and Technology programs. The assessment specialists' knowledge of the level of specificity needed to create the test blueprint (i.e., a document indicating the required numbers of items for each combination of content area, performance expectation, and item type).

In response to the question…	the team decided…	based on…
What types of test items should be included?	Multiple-choice and short-answer items.	The content experts' review of the types of items contributed by the colleges for possible inclusion in the assessment, their knowledge of the types of items typically used to assess the specified knowledge and skills, and their familiarity with the limitations of the computer interface for recording some types of responses. The assessment specialists' knowledge about the advantages and disadvantages of different types of items, in terms of cognitive demands, time required for students to respond, time and cost of scoring, contribution to test score reliability, etc.
How difficult should the items be?	The difficulty of the items should match the level of knowledge and skills required for the beginning-level college courses (approximately Grade 8 level).	The content experts' review of the curricula of beginning-level college courses. The assessment specialists' knowledge of how the difficulty of items can affect the motivation of students, the distribution of scores, and the precision of placement decisions.
For what subgroups of students should items be examined for possible differential difficulty (and, so, possible bias)?	Gender and first language.	The content experts' knowledge of the literature on mathematics performance. The assessment specialists' knowledge of the number of students required in each subgroup for the analyses.
Should each item measure a single or multiple content areas?	Each item should measure only one content area.	The content experts' knowledge of what kinds of subscores would be useful in making placement decisions and providing suggestions to students of areas for improvement. The assessment specialists' knowledge of the numbers of items and responses that would be needed for analyses
How many items in each area should be administered to each student?	At least four items per content area for the course placement version of the test; eight items per content area for the diagnostic version.	The content experts' knowledge of the minimum number of items that would be required to determine mastery of the important concepts in different content areas. The assessment specialists' evaluation of the numbers and types of items needed to generate subscores.
What should the cut scores be?	80% for acceptable performance and 60% for marginal performance in each content area.	The content experts' knowledge of typical item difficulty and the level of knowledge and skills needed for the entry-level college courses. The assessment specialists' knowledge of how to set cut scores.

In response to the question…	**the team decided…**	**based on…**
Should the test forms be used to make decisions about students' placement?	The forms will be used.	The content experts' review of the test content and format. The assessment specialists' review of the analysis results and other evidence supporting the validity argument.

Table 2. Validity argument for a mathematics course placement test

The response to the question…	leads to the validity argument inference…	which could be or has been supported by the following evidence…
What knowledge and/or skills should be tested?	The knowledge and skills selected by the test developers are the knowledge and skills students will need for success in college-level mathematics courses (ecological).	A review of the material being taught in the remedial courses found that it was more closely related to the mathematics curriculum at Grade 8 than to the high school mathematics curriculum.
What types of test items should be included?	The multiple-choice and short-answer items developed for the test measure the intended knowledge and skills (elemental).	A classical test theory (CTT) analysis of the responses of examinees in the field test to examine the item difficulty and internal consistency reliability (that is, that students who answered a particular item correctly also answered other items correctly).
How difficult should the items be?	The multiple-choice and short-answer items developed for the test are of the intended difficulty (elemental).	An item response theory (IRT) analysis to provide an estimate of item difficulty that is independent of the ability level of the small group of examinees who were administered each item in the field test.
For what subgroups of students should items be examined for possible differential difficulty (and, so, possible bias)?	The items are not differentially difficult (and so, potentially biased) by gender or first language (elemental).	Differential item functioning (DIF) analyses when sufficient data for the subgroups are available.
Should each item measure a single or multiple content areas?	Failure to answer an item correctly suggests weakness in the content area measured by that item (elemental).	A factor analysis to suggest whether performance on items that are intended to measure a content area is related to a single factor.
How many items in each area should be administered to each student?	The numbers of items will provide sufficiently reliable results for the intended uses (structural).	Based on the IRT results, calculation of the standard error of the ability estimates at different points on the ability scale for the proposed numbers of items in each content area.

The response to the question…	leads to the validity argument inference…	which could be or has been supported by the following evidence…
What should the cut scores be?	These cut scores represent the required levels of knowledge and skills for the intended uses (structural).	For each content area, plotting of the test characteristic curve (TCC) to translate between the IRT scale and the proportion correct scale; a formal standard setting.
Should the test forms be used to make decisions about students' placement?	There is sufficient evidence to support the intended uses (ecological).	Review of the evidence above; research on whether students who are recommended to take remediation are successful in subsequent mathematics courses that they take.

For each inference in Table 2, we have labelled it as elemental, structural, or ecological, according to Schilling and Hill's (2007) categories. As the table shows, we found that the elemental inferences did not necessarily precede the structural and ecological conclusions—for example, determining the content of the test (ecological) preceded deciding on the format of the items (elemental).

Discussion and Conclusions

It has been argued that much of the research on college placement test validation collects evidence in support of criterion validity alone, which has long been considered insufficient to support the validity of a test's use (AERA et al., 1999; Cronbach, 1988; Embretson, 2007; Kane, 2008; Lederman, 2011; Mislevy, 2007). Much of this evidence has been in the form of predictive validity studies, which examine how well placement tests can predict retention and graduation rates, given accurate placement (see Lederman, 2011, for a comprehensive review). Though this is relevant validity evidence, the validity argument approach makes it clear that other types of evidence are also needed. In addition, these studies do not consider the implications of decisions made during the development process for the validity of the score use; validity is considered only after the test has been administered (Ferrara, 2007). Further, in one of their critiques of Kane's validity argument approach, Schilling and Hill (2007) note the scarcity of real-world examples using an interpretive argument approach, speaking to a gap between validity theory and practice. Without systematically examining the tests' assumptions and inferences from the

beginning of test development, determining later whether test scores are being interpreted and used as intended is more difficult (Schilling, 2004; Schilling & Hill, 2007).

This article provides such an example of the use of a validity argument, while also showing how this approach can support collaborative test development, with a particular emphasis on the use of non-technical language. The validity argument made explicit why each decision is important and what types of evidence might inform or support each decision. It also made clear how the contributions of the content experts and assessment specialists were both essential to the quality of the resulting assessment. We look forward, in future collaborations, to continuing to refine this approach and extend it to other content areas and types of tests.

That the example is drawn from the Canadian context is relevant. Much of the literature on placement-test validation occurs in a US post-secondary context (Marwick, 2002; Medhanie, Dupuis, LeBeau, Harwell, & Post, 2012; Melguizo, Kosiewicz, Prather, & Bos, 2014; Ngo & Melguizo, 2016). To our knowledge, this is the first application of Schilling and Hill's (2007) validity argument approach in a Canadian college context.

It is important to note that the approach described here is principally focused on the initial development of an assessment; future work includes adapting it to accommodate the maintenance and revision of the assessment over time. This assessment will need to be periodically revised to make sure that the content areas remain relevant through changes in the secondary school mathematics and college program curricula. The current version of the assessment focuses on the mathematics used in college Business and Technology programs, but it may make sense to expand to other programs as well. As the assessment is used more widely and the items are seen by more students, it will also be important to replenish the items. Other areas for future development of this approach include making explicit the processes and rationale for reviewing new items, for sensitivity and possible cultural bias, for field testing new items, and for reviewing the field test results.

Finally, writing this article, like developing the mathematics assessment, was a collaboration between content experts and assessment specialists. Writing it made clear to us how much we have learned from each other and how our contributions combine to create something neither group could do alone. Based on our experiences, we recommend this approach, especially for teams that include members with complementary strengths, but who do not all have experience developing high-stakes assessments.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bahr, P. R. (2008). Does mathematics remediation work? A comparative analysis of academic attainment among community college students. *Research in Higher Education*, *49*(5), 420–450. https://doi.org/10.1007/s11162-008-9089-4

Bahr, P. R. (2013). The aftermath of remedial math: Investigating the low rate of certificate completion among remedial math students. *Research in Higher Education*, *54*(2), 171–200. https://doi.org/10.1007/s11162-012-9281-4

Bailey, T. (2009). Challenge and opportunity: Rethinking the role and function of developmental education in community college. *New Directions for Community Colleges*, *2009*(145), 11–30. https://doi.org/10.1002/cc.352

Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, *29*, 255–270. https://doi.org/10.1016/j.econedurev.2009.09.002

Ball, D., Bass, H., Boerst, T., Lewis, J., Sleep, L., Suzuka, K., & Thames, M. (2007, March). *Learning mathematical knowledge for teaching: Opportunities grounded in practice*. Presentation made at the annual meeting of the National Council of Teachers of Mathematics, Atlanta, GA.

Cohen, A. M., & Brawer, F. B. (2003). *The American community college*. San Francisco, CA: John Wiley.

College Student Achievement Project (CSAP) Team. (2015). *College student achievement project: Final report 2015* (prepared for the Ontario Ministry of Education and the Ontario Ministry of Training, Colleges and Universities). Toronto, ON: Sheridan College. Retrieved from http://csap.senecacollege.ca/docs/CSAP%20Cycle%202%20final%20report%2011Jun15.pdf

Cronbach, L. J. (1988). Five perspectives on validity. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*, 449–455.

Ferrara, S. (2007). Our field needs a framework to guide development of validity research agendas and identification of validity research questions and threats to validity. *Measurement: Interdisciplinary Research and Perspectives*, *5*(2–3), 156–164. https://doi.org/10.1080/15366360701487500

Fields, R., & Parsad, B. (2012). *Tests and cut scores used for student placement in postsecondary education, Fall 2011*. Washington, DC: National Assessment Governing Board.

Gal, I., & D. Tout. (2014). *Comparison of PIAAC and PISA frameworks for numeracy and mathematical literacy* (OECD Education Working Papers, No. 102). Paris, France: OECD Publishing. https://doi.org/10.1787/5jz3wl63cs6f-en

Hendrickson, A., Ewing, M., Kaliski, P., & Huff, K. (2013). Evidence-centered design: Recommendations for implementation and practice. *Journal of Applied Testing Technology*, *14*(1), 1–27.

Hoyles, C., Wolf, A., Molyneux-Hodgson, S., & Kent, P. (2002). *Mathematical skills in the workplace*. London, England: Science, Technology and Mathematics Council. Retrieved from http://discovery.ucl.ac.uk/id/eprint/10001565

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher*, *37*(2), 76–82. https://doi.org/10.3102/0013189x08315390

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice*, *23*, 309–311. https://doi.org/10.1080/0969594X.2016.1156645

Lederman, J. (2011). *Critical third-space phenomenology as a framework for validating college composition placement* (Doctoral dissertation). Indiana University of Pennsylvania, Indiana, PA. Retrieved from https://knowledge.library.iup.edu/etd/966

Marwick, J. D. (2002). *Charting a path to success: How alternative methods of mathematics placement impact the academic success of community college* (Doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana, IL. Retrieved from http://hdl.handle.net/2142/79700

Medhanie, A. G., Dupuis, D. N., LeBeau, B., Harwell, M. R., & Post, T. R. (2012). The role of ACCUPLACER mathematics placement test on a student's first college mathematics course. *Educational and Psychological Measurement*, *72*(2), 332–351. https/doi.org/10.1177/0013164411417620

Melguizo, T., Kosiewicz, H., Prather, G., & Bos, H. (2014). How are community college students assessed and placed in developmental math? Grounding our understanding in reality. *Journal of Higher Education*, *85*, 691–722. https://doi.org/10.1353/jhe.2014.0025

Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*(8), 463–469. https://doi.org/10.3102/0013189x07311660

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). *On the structure of educational assessments* (CSE Tech. Rep. 597). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://cresst.org/wp-content/uploads/TR597.pdf

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. Thousand Oaks, CA: Sage.

Ngo, R., & Melguizo, T. (2016). How can placement policy improve math remediation outcomes? Evidence from experimentation in community colleges. *Educational Evaluation and Policy Analysis*, *38*(1), 171–196. https://doi.org/10.3102/0162373715603504

Organisation for Economic Co-operation and Development (OECD). (2013). *Technical report of the Survey of Adult Skills (PIAAC)*. Paris, France: OECD Publishing.

Retrieved from https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

Orpwood, G., & Brown, E. (2015, April). *College Student Achievement Project Assessment Development Project: Final report* (prepared for the Ontario Ministry of Education and the Ontario Ministry of Training, Colleges and Universities). Retrieved from http://csap.senecacollege.ca/en/publications.php

Schilling, S. G. (2004). Conceptualizing the validity argument: An alternative approach. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 178–182.

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement: Interdisciplinary Research & Perspectives*, *5*(2/3), 70–80. https://doi.org/10.1080/15366360701486965

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–355). Washington, DC: American Council on Education

Scott-Clayton, J., Crosta, P., & Belfield, C. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis*, *36*, 371–393. https://doi.org/10.3102/0162373713517935

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Steen, L. A. (2001). *Mathematics and democracy: The case for quantitative literacy*. Washington, DC: Woodrow Wilson National Fellowship Foundation.