Title: Determining Mathematical Item Characteristics Corresponding With Item Response Theory Item Information Curves

Authors: Jim Gleason, Calli Holaway, and Andrew Hamric

Preliminary Research Report

Abstract:

Tests in undergraduate mathematics courses are generally high stakes, and yet have low reliability. The current study aims to increase the reliability of such exams by studying the qualities of test items that determine the ability of the item to contribute to the information of the test. Using a three parameter item response theory model, 695 items contained in 25 different tests for 5 different first-year undergraduate mathematics courses have been analyzed to determine the ability of each item to contribute to the corresponding test's reliability. During the conference presentation, the speakers with solicit input from the participants regarding the types of qualities of these items that may contribute to their information index. These qualities may include cognitive, mathematical content, linguistic, or other descriptions.

Keywords: Assessment, test writing, item response theory

Tests comprise a major component of mathematics classes at the undergraduate level, particularly first and second year courses. The grades on tests range from 30% clear up to 100% of a student's final grade. However, very little is known about the reliability of such tests that can dictate whether students pass or fail a course, or can cause a student to need an additional year to complete college, adding thousands of dollars to the student's college expenses. Through an analysis of final exams in College Algebra and Business Calculus using a three-parameter item response theory (IRT) model for 1438 and 524 students, respectively, we have found that for a student receiving the border-line score to advance to the next course of 70%, the standard error is between 10% and 14%. In other words, the student's actual score is somewhere between an F and a B when taking into account measurement error. This type of reliability is unacceptable for such a high stakes exam. The goal of this current research program is to determine the characteristics of test items that contribute the most to improving this reliability.

There are several ways to test a student's knowledge of a particular subject, with multiple-choice and constructed response the two most popular. Constructed response items include any assessment where the test taker does not have a list of formulated responses from which to choose. These types of questions require more resources to administer and grade than multiple-choice with a constructed response test of equivalent reliability to a multiple-choice test taking from 4 to 40 times as long to administer and is typically thousands of times more expensive (Wainer & Thissen, 1993; Lukhele, Thissen, & Wainer, 1994). However, with the rise of homework response systems, this difference in administration and grading is becoming negligible. Our analysis includes both constructed response and multiple choice items used on tests in Remedial Mathematics, Intermediate Algebra, Finite Mathematics, College Algebra, and Business Calculus.

This study uses 695 items from twenty-five tests from five different first year mathematics courses to determine what characteristics contribute the most to the item providing information contributing to a test's reliability. Of these items, 18% were constructed response, 3% were true/false or yes/no items, and the remaining 79% were multiple-choice. For each test, a three-parameter IRT model (van der Linden & Hambleton, 1997, pp. 13-17) was used to determine the appropriate difficulty and discrimination parameters for the items, with the guessing parameter fixed at 0 for constructed response items, 0.25 for multiple choice items with four choices, and 0.5 for dichotomous response items. Using the parameters generated from the model, the item information function for each item was multiplied by the student ability distribution function for the corresponding test and then integrated over the range of student abilities to generate an item information index.

The item information indices ranged from essentially zero to 5.948, with a mean of 0.332, a standard deviation of 0.397, and a median of 0.251. Additionally, 60% of the items had an information index less than the mean, implying that less that 40%% of the items contributed nearly all of the reliability for each test. If instructors could know which attributes contributed to items having a high item information index, then more mathematics tests would have the reliability appropriate for such high-stakes testing.

During the presentation, we will study the items with high item information indices while answering the following questions.

1. What are the cognitive categories that might be contributing to an item’s high information index?

   These cognitive categories could be based upon the structure of the observed learning outcome (SOLO) taxonomy (Biggs & Collis, 1982), Bloom’s taxonomy (Engelhart, Furst, Hill, & Krathwohl, 1956), or the mathematical tasks framework (Stein & Smith, 1998). The challenge is that these taxonomies were designed for situations other than analyzing test items, with some of the taxonomies shown to actually be ineffective in accurately categorizing items to predict student cognitive processes as they work on such items (Chan, Tsui, Chan, & Hong, 2002; Gierl, 1997). However, this does not exclude them from possible effectiveness in the current context.

2. What are the content oriented categories that might be contributing to an item’s high information index?

   While the goal of the current project is to discover ways of analyzing items that are independent of the mathematical topics assessed, there may be content oriented categorizations which contribute to an item’s information index. One such possible example is rational expressions. Student’s regularly have difficulty with fractions (Brown & Quinn, 2006), which may cause them to shut down when encountering rational expressions on a test and so may not perform as expected on such items even if the main goal of the item is to measure mathematical task distinct from rational expressions. On the other hand, such difficulty may contribute to the ability to differentiate students of various ability levels. Other similar topics may also exist and will be discussed among the participants.

3. What are the linguistic descriptors that might be contributing to an item’s high information index?

   Translating between mathematical language, visual information, and descriptive language is challenging for many students (Arcavi, 2003; Capraro & Joffrion, 2006; Radford & Puig, 2007). This challenge may contribute to the ability to distinguish between top students and weaker students and so may contribute to an item's information index.

4. Are there other constructs or lenses through which the test items may be analyzed?

Other constructs exist that the researchers have not thought about and will be sought from the participants in the conference presentation.

While the line of research proposed for this conference presentation is very undeveloped, it is an area with great promise due to the increase in information provided by the use of computerized assessment systems used in large settings and has the potential to greatly influence the future of classroom assessment in the college mathematics classroom.

Works Cited

Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics , 52* (3), 215-241.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: the SOLO taxonomy.* New York: Academic Press.

Brown, G., & Quinn, R. J. (2006). Algebra students' difficulty with fractions: An error analysis. *The Australian Mathematics Teacher , 62* (4), 28-40.

Capraro, M. M., & Joffrion, H. (2006). Algebraic equations: Can middle-school students meaningfully translate from words to mathematical symbols? *Reading Psychology , 27*, 147-164.

Chan, C. C., Tsui, M. S., Chan, M. Y., & Hong, J. H. (2002). Applying the structure of the observed learning outcomes (solo) taxonomy on students' learning outcomes: an empirical study. *Assessment & Evaluation in Higher Education , 27* (6), 511-527.

Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals (Handbook 1: Cognitive Domain).* (B. S. Bloom, Ed.) New York: Longman.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy. *The Journal of Educational Research , 91* (1), 26-32.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement , 31* (3), 234-250.

Radford, L., & Puig, L. (2007). Syntax and meaning as sensuous, visual, historical forms of algebraic thinking. *Educational Studies in Mathematics , 66*, 145-164.

Stein, M. K., & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics Teaching in the Middle School , 3*, 268-275.

van der Linden, W. J., & Hambleton, R. K. (1997). Item Response Theory: Brief History, Common Models, and Extensions. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 1-31). New York: Springer-Verlag.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education , 6*, 103-118.