Using Item Response Theory to Improve Locally-Constructed Multiple Choice Tests: Measuring Knowledge Gains and Curricular Effectiveness

Janie L. Knell
Morehead State University, USA

Andrea P. Wilhoite
Morehead State University, USA

Joshua Z. Fugate
Morehead State University, USA

Wilson J. González-Espada
Morehead State University, USA

Abstract

Current science education reform efforts emphasize teaching K-12 science using hands-on, inquiry activities. For maximum learning and probability of implementation among inservice teachers, these strategies must be modeled in college science courses for preservice teachers. About a decade ago, Morehead State University revised their science content courses to follow an inquiry approach. As part of the courses' assessment, a locally-made, diagnostic pre- and post-test was prepared. The main purpose of this "ex post facto" study was to demonstrate how concepts from Item Response Theory can be used to detect and remove psychometrically faulty items, and how the remaining items can be used by teachers to determine science learning gains in an inquiry-based physical science course that implemented two different curricula, "Physics and Everyday Thinking" and "Interactions in Physical Science".

Key words: Item response theory, assessment, physical science psychometrics

Correspondence should be addressed to: Dr. Wilson J. González-Espada, Department of Mathematics and Physics, Morehead State University, 405A Lappin Hall, Morehead, KY 40351. w.gonzalez-espada@moreheadstate.edu.

Introduction

Inquiry physical science curricula

Recently, results from the Program for International Student Assessment (PISA) suggested that, in the United States, school student performance in science

and mathematics has moved from world-class to middle-of-the-pack (Snyder & Dillow, 2013). Teacher academic preparation and quality has been pointed out as one factor that must be improved for PISA scores to recover. In the last decades, science education researchers have reported that students tend to learn better when science courses are interactive, collaborative, and inquiry-based. As a consequence, educators, particularly those who train preservice school teachers, should move away from more traditional, passive, memorization-oriented courses (Beiswenger, Stepans, & McClurg, 1998; Briscoe & Prayaga, 2004; Krockover, Shepardson, Eichinger, Nakhleh, & Adams, 2002; Luera & Otto, 2005; National Research Council, 2000, 2001).

At Morehead State University, a regional public university located in Eastern Kentucky, the transition to inquiry-based courses occurred around the year 2007. Before then, preservice elementary students were required to complete two lecture-based courses, Introduction to Physical Science and Introduction to Life Sciences.

The revised course sequence, informed originally by the National Science Education Standards (National Research Council, 1996) and currently by the Next Generation Science Standards (NGSS Lead States, 2013), resulted in three activity-based courses: Inquiry Physical Science (covering properties of matter, force and motion, heat, light and optics, electricity and magnetism, engineering and sound), Inquiry Earth and Space Science (covering astronomy, geology and meteorology), and Inquiry Life Sciences (covering cell structure and function, photosynthesis, respiration, reproduction, growth, heredity, evolution and ecology). This article reported data from one of these revised courses, Inquiry Physical Science (SCI 111).

Between 2008 and the present, SCI 111 has been taught by the same instructor, usually one or two sessions per semester. In addition to using several formative and summative assessments through the semester (including daily quizzes, unit tests, video reports, written projects, and homework), the course was evaluated using a locally-made, diagnostic assessment. This test consisted of 40 questions directly correlated with the content of the course. Although many test questions were inspired by other validated assessments, the final diagnostic test only went through face validity by a panel of experts. This is true for many locally-made, classroom tests used in schools and postsecondary education institutions.

Between 2008 and 2010, SCI 111 used a research-based curriculum called "Physics for Everyday Thinking" (PET), created by *Its About Time, Inc.* (Goldberg, Robinson, & Otero, 2008). They described PET as follows (It's About Time Inc., 2015a):

> *PET is a one-semester curriculum designed in part for prospective or practicing elementary teachers. The course uses a student-oriented pedagogy with a physics content focus as well as a unique Learning about Learning component. It has been taught at two-year and four-year institutions; has been adapted for a science methods course in schools of education; and can be offered as a*

> *workshop for practicing elementary teachers. PET elicits student initial ideas and then provides students with opportunities to acquire evidentiary support, through hands-on activities or computer simulations, which helps them to decide, if appropriate, to develop new or modified ideas. This component is designed to help students develop an understanding of how scientists develop knowledge, how they learn science themselves, and how others (for example, either elementary school students or other college students) learn science.*

Between 2010 and the present, a new research-based curriculum called "Interactions in Physical Science" (IPS) was implemented to cover additional properties of matter and modern physics topics (Goldberg, 2009). IPS was created by the same company, and described this course as follows (It's About Time Inc., 2015b):

> *The content in IPS is broken down into carefully crafted chapters of learning. Each chapter begins with a purpose followed by a Key Question. Students generate ideas and questions, then explore using the science practices. They record their results and, like scientists, they discuss their results with each other and as a class. Students also compare their ideas with real scientists. The role of eliciting students' prior knowledge is an important aspect of the pedagogy of IPS. The appreciation of the importance of students' initial ideas, as well as the need to reconcile those ideas with formal learning, guided the development of the curriculum. IPS is hierarchical, in that chapters and units build on one another, and social, because real scientific knowledge develops through collaboration as communities of scientists work together. In IPS, students, like scientists, interact with their peers as they work in teams to do experiments and gather evidence, share ideas with their group, and participate in class discussions to build consensus ideas.*

Table 1 summarizes the main content included in each of these curricula.

Physics and Everyday Thinking	Interactions in Physical Science
* Interactions and energy - speed; motion and energy; contact interactions; slowing and stopping; warming and cooling; conservation of energy. * Interactions and forces - motion with a continuous force; pushes and slowing down; net force; friction; Newton's laws; motion with balanced forces. * Interactions and systems - magnetic interactions; electric charge interactions; gravitational interactions. * Model of magnetism -	* Science experiments – measurements; experimental design. * Introducing interactions – magnetism; electric charge; electric circuits; electro-magnetism. * Interactions and properties – measuring length volume and density; characteristic properties. * Energy descriptions of interactions – energy; mechanical waves; energy transfer; speed of objects and waves; changing speed. * Mechanical interactions – motion energy; applied, friction, drag and elastic interactions. * Mechanical interactions and forces – forces; frictionless motion; net force; Newton's laws; simple machines. * Gravitational interactions – Law of gravitation; mass and weight; orbital motion; terminal speed; buoyancy; potential

experiments with magnets; initial and improved experimental models. * Electric circuit interactions - circuits and energy; multi-bulb circuits and energy; multi-bulb circuits and current; electrical efficiency. * Light interactions – shiny surfaces, light and vision; non-shiny and black surfaces; refraction; light and color.	energy. * Mass conservation in open and closed systems. * Energy conservation – Heat conduction and infrared interactions; thermal energy and phase change; efficiency; reflection and refraction; color. * Chemical interactions – Acids and bases; burning reactions; exothermic and endothermic reactions. * Interactions and classifying materials – physical interactions; mixture and single substance; elements and compounds; the Periodic Table. * Physical interaction and the structure of materials – properties of solids, liquids and gases; atomic structure; isotopes and radioactivity. * Chemical reactions – Chemical bonds, balancing reactions.

Since the same locally-made diagnostic assessment was offered since 2008, enough students have completed it, making it possible to measure with a high degree of confidence both the students' science knowledge gains and to compare what curricula, PET or IPS, better contributed to these gains.

In this case, the sample size was particularly important because it allowed the use of psychometric techniques to evaluate the quality of individual test items, to identify those items that have questionable psychometric parameters, and to remove such items prior to any knowledge gains analysis. This process strengthened the confidence in the study's findings.

Item Response Theory and Locally-Made Assessments

Assessments are commonly used in the fields of education and psychology (Alkharusi, Aldhafri, & Alnabhani, 2014; Jones, 2013; Moss, Girard, & Haniford, 2006). One of the greatest issues in education has been to determine how to measure learning and other instructional constructs that cannot be physically measured (Ardovino, 2000; Sternberg, 2003). Despite this problem, educators measure their students' aptitude and learning through various forms of assessments, especially written tests (Chatterji, 2003).

What many teachers have failed to recognize is that, just because students missed a test question, it might not necessarily mean that the students failed to learn the material. Sometimes the test itself was not rigorously constructed or might have validity and reliability problems (Brown, 2000; Koretz, 2008; Lemann, 1999; Weller, 2001).

Fortunately, a number of psychometric approaches have been developed to enhance the reliability and validity of measures, under the theoretical framework of Classical Test Theory (de Klerk, 2014; DeVellis, 2006; Mislevy, 1996). These types of analyses focus on a whole test rather than specific items on a test (Hambleton & Jones, 1993; Lord, 1959; Zimmerman, 1998).

A more recent approach to enhance the validity and reliability of tests is Item Response Theory or IRT (Morris et al, 2006). Where classical test theory measured results of a test as a whole, IRT can analyze responses to individual items on a test (Erdodi, 2012; Fan, 1998; Gonzalez-Espada, 2008, 2009; Hill & Lewicky, 2007; Pellegrino, 2001; Yang, 2014). IRT experts have suggested that test items that are answered correctly by almost everybody, items that almost no one answered correctly, and items where below-average scorers performed better than above-average scorers should not be part of many assessments (Crocker & Algina, 1986).

Despite all the available information on IRT, this topic is seldom covered in many pre-service education measurement courses and, as a consequence, it is rarely used in the classroom on teacher-made tests (Morris et al, 2006). Only large-scale standardized tests, such as the ACT and GRE tests, have faced the scrutiny of item analysis (Pellegrino, 2001; Wagner, 2008). Fortunately, computer technology and spreadsheets are ubiquitous now, allowing teachers to start computing data-rich IRT parameters for locally-made tests, especially diagnostic tests, end-of-course tests, or unit tests that are used over subsequent semesters, where appropriate sample sizes can be obtained.

Although many standardized assessments have been through IRT validation prior to being implemented, this has not always been possible for locally-made tests. As a consequence, an ex post facto IRT validation approach can be an option. In this case, students' responses are used to flag problematic items and to remove them from score calculations; the remaining items can then provide a more robust dataset for instructional and evaluation purposes.

Purpose

This study had two main goals: (1) to use IRT concepts to demonstrate how a locally-made diagnostic test can be evaluated "ex post facto" to identify questions that do not meet psychometric parameters and remove them prior to an analysis of knowledge gains, and (2) to use diagnostic test data to determine what inquiry-based physical science curricula, PET or IPS, resulted in the largest science knowledge gain among pre-service students enrolled in SCI 111 at Morehead State University.

Methods

Sample

The population of interest was made of college students majoring in elementary education (P-5) at Morehead State University. The sample size was 456 students (86% female; 12% male) who enrolled in SCI 111, including 35% freshmen, 33% sophomores, 24% juniors, and 7% seniors. Of these students, 278 used the PET curriculum between 2008-2010 and 175 used the IPS curriculum between 2010-2013. A total of 192 students took both the pre- and the post-test.

Diagnostic Test

The diagnostic test for SCI 111 consisted of 40 multiple-choice items with five alternatives. The content emphasis of this assessment was physical science, including topics such as properties of matter, linear and circular motion, forces, light and sound waves, heat and temperature, magnetism, and electricity and circuits. This test was prepared by combining items from different sources, including previously validated instruments, such as the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992; Well, Henstenes, & Swackhamer, 1995). The pre-test was completed the first day of class, after going over the class syllabus. The post-test was completed the last week of the semester, before the end-of-course examination. The semester was 15 weeks long.

Statistical Analysis for IRT Parameter Calculation

Before any science content gains can be calculated, two important IRT parameters were calculated, item difficulty and item discrimination. These were compared with the suggested values psychometricians consider appropriate. Those items that do not meet IRT guidelines were discarded from the dataset.

For each item, the ratio of correct scores and the sample size for a given item was computed. This is known as the item difficulty:

$$\text{Difficulty} = \frac{\text{correct scores}}{\text{sample size for item}}$$

The literature suggested that items that are answered incorrectly by most students (difficulty < 0.20) or items that are answered correctly by most students (difficulty > 0.80) diminish the validity and reliability of the test as a whole and should be discarded (Crocker & Algina, 1986; Ebel & Frisbie, 1991).

In addition, overall scores were used to classify students into two sub-groups, "above average scorers" and "below average scorers". Then, item difficulty was calculated for each item and each subgroup. Finally, the difference between the item difficulties of the subgroups, known as the item discrimination, was calculated:

$$\text{Discrimination} = \text{difficulty}_{\text{above average scorers}} - \text{difficulty}_{\text{below average scorers}}$$

Basically, item discrimination measured the degree to which students with high overall exam scores also answered a particular question correctly. A question was considered a good discriminator when students who answered the question correctly also did well on the test (Slater, Beal-Hodges, & Reed, 2014). The literature suggested that the discrimination of an item should be positive, that is, above average scorers should do better on an item compared with below average scorers. In general, item discrimination values between 0.4 and 1.00 are considered best, and values between 0.20 and 0.40 are satisfactory. When an item discrimination value is below 0.20, it is not differentiating well between low and high scorers. An item with discrimination values close to zero or negative must be discarded (Crocker & Algina, 1986; Ebel & Frisbie, 1991).

Statistical Analysis for Knowledge Gains Calculations

After removing psychometrically problematic items from the dataset, the rest of the items, and the overall scores, were analyzed using standard descriptive and inferential statistics (Weinberg & Goldberg, 1990). The average pre-test scores for both PET and IPS, the average post-test scores for both PET and IPS, the average pre- and post-test scores for PET, and the average pre- and post-test scores for IPS were compared using t-tests to identify significant differences. A p-value of 0.05 was selected as a cutoff value to balance the possibility of both Type I and II errors.

In addition, normalized gains, also known as Hake gains (Hake, 1998), were calculated for each test item. This formula established a ratio between the number of correct answers in the pre- and post- surveys for any given item and the difference between the maximum possible score and the pre-survey score for that item. Since 192 students completed both the pre- and the post-test, the formula becomes:

$$\text{Normalized gain} = \frac{[\text{post-survey item score}] - [\text{pre-survey item score}]}{192 - [\text{pre-survey item score}]}$$

A normalized gain factor indicated growth in the construct of interest with respect to the participants' starting position, mathematically reducing potential ceiling effects if the scores are close to 100%. The literature (Hake, 1998) established standard cutoff points as follows: A normalized gain of less than 30% was considered "low", one between 30% and 70% was considered "moderate", and one above 70% was considered as "high gain".

Findings and Discussion

IRT Post-validation Analysis

After removing from the dataset students who completed the pretest but not the posttest (n = 63, 13.8% of the total number of participants) and students who completed the posttest but not the pretest (n = 6, 1.3% of the total number of participants), the dataset was reduced to 192 students who completed both tests. The revised dataset was composed of 46 male students (12%) and 332 female students (88%) distributed by class rank as 126 freshmen (34%), 128 sophomores (34%), 89 juniors (24%) and 29 seniors (8%). Six students did not identify their gender and 12 students did not identify their class rank. A total of 252 students (66%) completed SCI 111 using the PET curriculum and the remaining 132 students (34%) completed SCI 111 using the IPS curriculum.

The revised dataset was used to calculate item difficulty and item discrimination parameters. The results are shown in Table 2.

Question	Difficulty Pretest	Difficulty Posttest	Discrimination Pretest	Discrimination Posttest	Flagged
1	0.4219	0.5365	0.1114	0.0949	
2	0.8229	0.8333	0.0968	0.1230	x
3	0.8333	0.8906	0.0764	0.0229	x

4	0.3698	0.5938	0.2551	0.2028	
5	0.4896	0.5340	0.1871	0.2778	
6	0.7969	0.9010	0.0645	0.0840	x
7	0.2461	0.2604	0.1858	0.1991	
8	0.5260	0.8750	0.1157	0.0044	
9	0.1667	0.4844	0.0486	0.2562	
10	0.3646	0.7604	0.1819	0.0798	
11	0.6979	0.7552	0.2167	0.1324	
12	0.8958	0.9119	-0.0461	0.0205	x
13	0.1771	0.4583	0.1116	0.3846	
14	0.4948	0.6667	0.3646	0.4231	
15	0.6302	0.7292	0.1826	0.1665	
16	0.0628	0.0781	0.0858	0.0223	x
17	0.4688	0.6042	0.0195	0.0242	x
18	0.4031	0.7917	0.1907	0.2144	
19	0.7083	0.8438	0.1545	0.1634	
20	0.224	0.5104	0.1448	0.1083	
21	0.5885	0.7053	0.1809	0.2131	
22	0.4271	0.5469	0.2263	0.1561	
23	0.7906	0.8125	0.1401	0.1779	
24	0.1042	0.4427	-0.0165	0.2719	
25	0.1146	0.3802	0.1090	0.3206	
26	0.2917	0.3698	0.2831	0.2802	
27	0.1979	0.2448	0.0499	0.2320	
28	0.2135	0.4115	0.2486	0.4002	
29	0.1152	0.0573	0.1294	-0.0169	x
30	0.3298	0.5288	0.2512	0.1745	
31	0.7958	0.9271	0.2554	0.1233	
32	0.2513	0.3906	0.1965	0.2363	
33	0.5079	0.7083	0.1102	0.1579	
34	0.4263	0.5469	0.3353	0.3641	
35	0.4869	0.8639	0.3400	0.1118	
36	0.1937	0.5104	0.0375	0.1290	
37	0.5812	0.7813	0.2382	0.0359	
38	0.3968	0.7760	0.2687	0.0385	
39	0.3979	0.7016	0.0544	0.2568	
40	0.3158	0.1390	-0.0776	0.0279	x

Table 2. Item difficulty and discrimination parameters for a locally-made, diagnostic assessment. Items that did not meet IRT guidelines were “flagged”.

IRT data for both test administrations uncovered that questions 2, 3, 6, and 12 had item difficulty values higher than about 0.80, suggesting that most students

answered correctly, and item discrimination values lower than about 0.10, suggesting that low and high scorers were answering in similar ways. These questions were discarded from the dataset.

Also, for both test administrations, questions 16 and 29 had item difficulty values lower than about 0.20, suggesting that most students did not answer correctly, and item discrimination values are less than about 0.10, suggesting that low and high scorers were answering in similar ways. These questions were discarded from the dataset as well.

Question 40 was discarded because students found it more difficult in the posttest, the first discrimination value was negative, and the second one was very close to zero.

Question 17 was discarded upon further inspection of the students' responses. Although the discrimination values are very close to zero, the difficulty values were located between 0.50 and 0.60, which is normally considered appropriate. After a detailed examination of how students selected each alternative, it was noted that three of the options were basically ignored, converting this item into a 2-option question. The proportion of correct answers was close enough to 50-50 to have a strong guessing effect. Many of the remaining items had either good to marginal difficulty values or good to marginal discrimination values, but not both, and were preserved for the subsequent knowledge gains analysis.

Note that items were discarded without trying to determine why item difficulties and discrimination values were below recommended guidelines. When assessments are pre-validated, flagged items can be examined and revised as needed. Post-validated assessments, like the one used in this study, do not have that advantage. At this point, it is simply impossible to guess how and why students responded to a flagged item, although incorrectly keyed answers, confusing text, confusing illustrations, content that was not thoroughly covered during class, or higher level questions might be partly responsible (Slater, Beal-Hodges, & Reed, 2014).

Science Knowledge Gains Analysis

Descriptive and inferential statistics were calculated with the remaining 32 diagnostic test items. Comparing the results of the pre- and post-test for the PET curricula, it was found that the average pre-score and standard deviation were 12.44 ± 3.51 and the average post-score and standard deviation were 19.86 ± 4.08. This difference was statistically significant ($t = 15.46$, $df = 250$, $p < 0.000$, effect size = 0.70). The results of the pre- and post-test for the IPS curricula were similarly compared, and it was found that the average pre-score and standard deviation were 14.00 ± 3.70 and the average post-score and standard deviation were 18.61 ± 4.03. This difference was also statistically significant ($t = 6.92$, $df = 132$, $p < 0.000$, effect size = 0.52). This means that both curricula performed similarly in producing statistically significant knowledge gains as assessed by the IRT-corrected diagnostic test.

However, it was noted that the actual range of average pre- and post-test scores for PET and IPS were 7.42 and 4.61 points respectively, suggesting that PET produced the largest increase in actual test points. The IPS curriculum covered more content knowledge than PET, especially in topics such as the nature of science, properties of matter, and modern physics. This, added to the inclusion of engineering topics required by the Next Generation Science Standards, might be causing an accelerated pacing that impacted the students' acquisition of content knowledge and resulted in a smaller pre- and post-test point range.

Comparing the results of the pretest for both curricula, it was found that the average pre-score and standard deviation for PET and IPS were 12.44 $\pm$ 3.51 and 14.00 $\pm$ 3.70, respectively. This difference was statistically significant ($t = 2.863$, $df = 190$, $p = 0.005$), suggesting that the 2008-2010 students, as a group, were statistically different from the 2010-2013 group of students. This result is intriguing because it would be expected that pre-test scores should be similar, regardless of when students enrolled in the class. A possible explanation for this result might be that, around 2011, the College of Education increased entrance requirements for their Teacher Education Programs, so it was possible that more recent students in SCI 111 had better grade point averages, which could be reflected in their prior science content knowledge.

The results of the posttest for both curricula were similarly compared, and it was found that the average post-score and standard deviation for PET and IPS were 19.86 $\pm$ 4.08 and 18.61 $\pm$ 4.03, respectively. An analysis of covariance demonstrated that the difference was also statistically significant ($F = 13.91$, $p < 0.000$).

For each test item, the normalized gain was calculated (Table 3). The overall gain was 0.35. Following Hake (1998), 15 items obtained a normalized gain of less than 30%, considered "low", 15 items obtained a normalized gain of between 30% - 70%, considered "moderate", and two items obtained a normalized gain above 70%, considered "high gain".

Question Number	Number Correct Pretest	Number Correct Posttest	Hake Gain	Question Number	Number Correct Pretest	Number Correct Posttest	Hake Gain
1	81	103	0.20	23	151	156	0.12
4	71	114	0.36	24	20	85	0.38
5	94	102	0.08	25	22	73	0.30
7	47	50	0.02	26	56	71	0.11
8	101	168	0.74	27	38	47	0.06
9	32	93	0.38	28	41	79	0.25
10	70	146	0.62	30	63	101	0.29
11	134	145	0.19	31	152	178	0.65
13	34	88	0.34	32	48	75	0.19

14	95	128	0.34	33	97	136	0.41
15	121	140	0.27	34	81	105	0.22
18	77	152	0.65	35	93	165	0.73
19	136	162	0.46	36	37	98	0.39
20	43	98	0.37	37	111	150	0.48
21	113	134	0.27	38	75	149	0.63
22	82	105	0.21	39	76	134	0.50

Table 3. For each question in the revised dataset, the number of students who answered correctly the pre- and post-test items is shown, as well as the normalized gain.

Conclusion and Limitations

One of the research study's goals was to use IRT concepts to demonstrate how a locally-made diagnostic test can be evaluated "ex post facto" to identify questions that do not meet psychometric parameters and remove them prior to an analysis of knowledge gains. Data analysis from almost 200 students was able to pinpoint eight questions that were either answered correctly by almost everybody, answered incorrectly by almost everybody, and/or questions where below-average scorers performed better than above-average scorers.

This finding was important because it shows that, even after using questions from previously validated assessments to create a diagnostic physical science test and after a panel of experts revised and approved it, the assessment still had questions that were psychometrically problematic upon further analysis using IRT. This situation might be similar to what happens in many science classrooms, where science teachers, as content experts, prepare tests with questions that look satisfactory (that is, the test has face validity), but might include questions that could be interpreted by students in different ways. This study demonstrated that, given a large enough sample size, science teachers could use IRT concepts to post-validate and improve their diagnostic, end-of-course, and unit assessments.

The second goal of this study was to use an IRT-improved dataset to compare the effectiveness of two physical science curricula, PET and IPS and which contributed to the largest science knowledge gain among students who completed SCI 111. The data showed that although both curricula resulted in statistically significant better scores on the post-test and an average normalized gain of about 35%, the PET curriculum produces a larger difference between average pre- and post-test scores. For teacher educators who are considering whether to implement PET or IPS in their inquiry physical science courses for preservice teachers, PET seems like a better option. This study also observed significantly higher pre-test scores among more recent students, which might be a reflection of increased entry requirements into the university's Teacher Education Program.

One main limitation of the study was sample size. Organizations that engage in creating, validating, and administering large-scale standardized tests can obtain very robust sample sizes for their IRT analyses, calculate very precise item parameters, and make better informed decisions about revising test questions prior to full implementation. Schoolteachers and college faculty will very likely not achieve large sample sizes, but even a moderate sample can lead to useful post-administration insights on test questions that might not meet IRT guidelines.

Overall, it was clear that even the most carefully prepared teacher tests need to be examined from an IRT perspective, especially unit tests or tests that are used in multiple semesters. In an age of increased accountability, teacher can learn from the results of this study to improve the validity, reliability, and accuracy of locally-made science assessments.

Acknowledgments

This research was supported by three Undergraduate Research Fellowships from the Department of Mathematics and Physics, College of Science and Technology, Morehead State University.

References

Alkharusi, H., Aldhafri, S., & Alnabhani, H. (2014). Classroom assessment: Teacher practices, student perceptions, and academic self-efficacy beliefs. *Social Behavior and Personality*, *42*(5), 835-855.

Ardovino, J. H. (2000). *Multiple measures: Accurate ways to assess student achievement*. Thousand Oaks, CA: Corwin Press.

Beiswenger, R. E., Stepans, J. I., & McClurg, P. A. (1998). Developing science courses for prospective elementary teachers. *Journal of College Science Teaching*, *27*(4), 253–257.

Briscoe, C., & Prayaga, C. S. (2004). Teaching future K-8 teachers the language of Newton: A case study of collaboration and change in university physics teaching. *Science Education*, *88*(6), 947–969.

Brown, J. (2000). *What is construct validity?* Shiken: JALT Testing & Evaluation SIG Newsletter, 4(2), 8-12. Retrieved from http://jalt.org/test/PDF/Brown8.pdf.

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston, MA: Allyn and Bacon.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich Publishers.

de Klerk, G. (2014). Classical test theory. In M. Born, C.D. Foxcroft & R. Butter (Eds.), *Online Readings in Testing and Assessment, International Test Commission*. Retrieved from http://www.intestcom.org/Publications/ORTA.php.

DeVellis, R. F. (2006). Classical test theory. *Medical Care: Measurement in a Multi-Ethnic Society*, *44*(11), S50-S59.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice Hall.

Erdodi, L. A. (2012). What makes a test difficult? Exploring the effect of item. *Journal of Instructional Psychology*, *39*(3-4), 171-176.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person parameters. *Educational and Psychological Measurement*, *58*(3), 357–381.

Goldberg, F. (2009). *Interactions in Physical Science, 2nd. Ed*. Mount Kisko, NY: It's About Time.

Goldberg, F., Robinson, S., & Otero, V. (2008). *Physics and Everyday Thinking*, 2nd. Ed. Mount Kisko, NY: It's About Time.

González-Espada, W. J. (2009). Detecting gender bias through test item analysis. *The Physics Teacher*, *47*(3), 175-179.

González-Espada, W. J. (2008). Physical science lab quizzes: Results from test item analysis. *Journal of Science Education/REC*, *9*(2), 81-85.

Hake, R. (1998). Interactive engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64-74.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141-158.

Hill, T. & Lewicki, P. (2007). *Statistics: Methods and applications*. Tulsa, OK: StatSoft Inc. Retrieved from http://www.statsoft.com/Textbook/Reliability-and-Item-Analysis.

It's About Time Inc. (2015a). *Physics and Everyday Thinking*. Retrieved from http://www.iat.com/courses/college-sciences/physics-and-everyday-thinking/

It's About Time Inc. (2015b). *Interactions in Physical Science*. Retrieved from http://www.iat.com/courses/middle-school-science/interactions-in-physical-science

Jones, R. (2013). Assessment and legal education: What is assessment, and what does it have to do with the challenges facing legal education? *McGeorge Law Review*, *45*(1), 85-110.

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Krockover, G. H., Shepardson, D. P., Eichinger, D., Nakhleh, M., & Adams, P. E. (2002). Reforming and assessing undergraduate science instruction using collaborative action-based research teams. *School Science and Mathematics*, *102*(6), 266–284.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus and Giroux.

Lord, F. M. (1959). Problems in mental test theory arising from errors of measurement. *Journal of the American Statistical Association*, *54*(286), 472-479.

Luera, G. R., & Otto, C. A. (2005). Development and evaluation of an inquiry-based elementary science teacher education program reflecting current reform movements. *Journal of Science Teacher Education*, *16*(3), 241–258.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*(4), 379-416.

Morris, G. A., Barnum-Martin, L., Harshman, N., Baker, S. D., Mazur, E., Dutta, S., Mzoughi, T. & McCauley, V. (2006). Testing the test: Item response curves and test quality. *American Journal of Physics*, *74*(5), 449-453.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity of educational assessment. *Review of Research in Education*, *30*(1), 109-162.

National Research Council. (2001). *Educating teachers of science, mathematics, and technology: New practices for the new millennium*. Washington, DC: National Academy Press.

National Research Council. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academy Press.

National Research Council (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: The National Academies Press.

Pellegrino, J. W. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Slater, R. D., Beal-Hodges, M., & Reed, A. (2014). Using Excel pivot table function for visual data analysis of exam results: A supplemental procedure to classical test theory. *Academy of Educational Leadership Journal*, *18*(4), 221-229.

Snyder, T. D. & Dillow, S. A. (2013). *Digest of education statistics*. Washington, DC: National Center for Educational Statistics.

Sternberg, R. J. (2003). *Models of intelligence: International perspectives*. Washington, DC: American Psychological Association.

Wagner, T. (2008). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need--and what we can do about it*. New York, NY: Basic Books.

Weinberg, S. & Goldberg, K. (1990). *Statistics for the Behavioral Sciences*. New York, NY: Cambridge University Press.

Weller, D. (2001). Building validity and reliability into classroom tests. *NASSP Bulletin*, *85*(622), 32-37.

Wells, M., Hestenes, D., & Swackhamer, G. (1995). A modeling method for high school physics instruction. *American Journal of Physics*, *63*(7), 606-619.

Yang, F. M. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171-177.

Zimmerman, D. W. (1998). How should classical test theory have defined validity? In B. D. Zumbo (Ed.), *Validity theory and the methods used in validation: Perspectives from social and behavioral sciences*. (pp. 233-251). Boston, MA: Kluwer Academic Publishers.