HOLTZCLAW, J. D.; EISEN, A.; WHITNEY, E. M.; PENUMETCHA, M.; HOEY, J. J. & KIMBRO, K. S. Incorporating a New Bioinformatics Component into Genetics at a Historically Black College: Outcomes and Lessons, *Cell Biology Education* **5**, 52-64, 2006.

HEERMAN, D. W. & FUHRMANN, T. T. Teaching physics in the virtual university: the mechanics toolkit, *Computer Physics Communications* **127**, 11-15, 2000.

HUGHES, I. E. Alternatives to laboratory practicals - do they meet the needs? *Innovations in Education and Teaching International* **38** (1), 3-7, 2000.

KEEVES, J. P. Methods and Processes in Research in Science Education. In: FRASER, B. J. and TOBIN, K. G. (eds.), *International Handbook of Science Education*, Part Two, Kluwer Academic Publishers, Dordrecht, Netherlands, 1127-1153, 1998.

MILLER, L.; MORENO, J.; WILLCOCKSON, I.; SMITH, D. & MAYES, J. An Online, Interactive Approach to Teaching Neuroscience, *Cell Biology Education* **5**, 137-143, 2006.

SCHARFENBERG, F. J.; BOGNER, F. X.; KLAUTKE, S. The Suitability of External Control-Groups for Empirical Control Purposes: a Cautionary Story in Science Education Research, *Electronic Journal of Science Education* **11** (1), 22-36, 2006.

SMITH, A. C.; STEWART, R.; SHIELDS, P.; HAYES-KLOSTERIDIS, J.; ROBINSON, P. & YUAN, R. Introductory Biology Courses: A Framework To Support Active Learning in Large Enrollment Introductory Science Courses, *Cell Biology Education* **4**, 143-156, 2005.

WILSON, C. D.; ANDERSON, C. W.; HEIDEMANN, M.; MERRILL, J. E.; MERRITTB, B. W.; RICHMOND, G.; SIBLEY, D. F. & PARKER, J. M. Assessing Students' Ability to Trace Matter in Dynamic Systems in Cell Biology, *Cell Biology Education* **5**, 323-331, 2006.

Received: 11.06.2007 / Approved: 20.04.2008

Physical science lab quizzes: results from test item analysis

Exámenes en laboratorios de física: análisis de los resultados de las preguntas de selección múltiple

WILSON J. GONZÁLEZ-ESPADA

School of Physical and Life Sciences, Arkansas Tech University, 1701 N. Boulder Ave. (McEver Hall, 203), Russellville, AR 72801, USA, wgonzalezespada@atu.edu

Abstract

Teachers have a difficult task when they assess students properly because they cannot directly measure mental constructs such as "knowledge" and "understanding". Teachers can use multiple choice items as a way to estimate student knowledge in a fast, inexpensive and reliable way, assuming that the items are properly designed and validated. Test item analysis borrows from large-scale test theory and can reveal significant facts about a classroom test, including technical flaws and errors of judgment made by the item writer, multiple interpretations of ambiguous items, poor distractors, and student misconceptions. This paper applies the concepts of item difficulty and discrimination in the context of the analysis of lab quizzes offered to more than 100 students enrolled in the "Introduction to Physical Science" course at Arkansas Tech University. The author found that most of the test items were easier than expected but with reasonable and high discrimination. However, several items were flagged as too easy or too difficult. Given their marginal level of discrimination, these items should be further analyzed for possible modification.

Key words*: assessment, testing, multiple-choice items, difficulty, discrimination, physical science*

Resumen

Dada la naturaleza abstracta de los constructos "conocimiento" y "entendimiento", evaluar directamente el aprendizaje de los estudiantes es difícil. Los ítems de opción múltiple son una manera rápida, accesible y confiable de estimar cuánto los estudiantes aprendieron en clase, pero sólo si se redactan de manera válida y confiable. El análisis de los exámenes por el maestro, utilizando algunas técnicas comúnmente aplicadas a las pruebas estandarizadas, puede revelar problemas con los ítems, tales como ambigüedad, errores de juicio del que redacta el ítem y distractores de poca calidad. También puede revelar aspectos positivos, tales como concepciones erróneas de los estudiantes. El propósito de este artículo es aplicar los conceptos de dificultad y discriminación al análisis de varios exámenes de selección múltiple completados por más de 100 estudiantes matriculados en el curso Introducción a las Ciencias Físicas en Arkansas Tech University. Se descubrió que muchos de los ítems tenían poca dificultad y mediana-alta discriminación. También se observó que algunos ítems eran muy fáciles o muy difíciles y de baja discriminación, por lo cual se examinarán y revisarán posteriormente.

Palabras clave*: evaluación, preguntas, selección múltiple, dificultad, discriminación, ciencias físicas.*

INTRODUCTION

The college faculty have the ineludible task of assessing students, which is one of the most difficult tasks because mental constructs cannot be measured directly. In fact, many publications address the theoretical foundations of assessment, the best ways to measure student learning, and the limitations of different types of assessments (CROCKER & ALGINA, 1986; HOGAN, 2007; JOHNSTONE & AMBUSADAI, 2001; NITKO, 1996; RACE, 2003; THORNDIKE, ANGOFF & LINDQUIST, 1971). A subgroup of these, focalise on science assessment (MINTZES, WANDERSEE, & NOVAK, 2000; ENGER & YAGER, 2001; HEDGES, 1966). Recently, many physics education researchers have turned their attention to assessment (DANCY & BEICHNER, 2006; HAZEL, LOGAN & GALLAGHER, 1997; SLATER, RYAN & SAMSON, 1997; O'BRIEN-PRIDE, VOKOS & MCDERMOTT, 1998; THORNTON & SOKOLOFF, 1998).

According to EBEL & FRISBIE (1986), tests as a whole can be assessed for a number of characteristics, including:

1. *Relevance*: Is the test a reflection of the content that was covered in class?
2. *Balance*: Does the test contain a weighted sample of all the important knowledge, skills, and understandings covered based on teacher emphasis in class?
3. *Efficiency*: Does the test yield a large number of independently scorable responses per unit of testing time?
4. *Specificity*: Is the test score near chance levels for a person not familiar with the subject matter?
5. *Difficulty*: Does the test have manageable difficulty levels?
6. *Discrimination*: How good is the test in identifying students with different levels of subject matter knowledge?
7. *Validity*: Does the test measure what it is intended to measure?
8. *Reliability*: Will students with the same level of subject matter knowledge, obtain about the same score on the test?

Multiple choice items are one of the most common ways to assess student knowledge in a fast and an inexpensive way. If instructors properly design and validate them, the multiple choice items can yield much information about the students' physics knowledge. Instructors may find a problem with the use of multiple choice items on a class test because they might not have the proper pedagogical content knowledge (SHULMAN, 1986) to prepare them, especially in how to write clear and concise stems, one unequivocally correct answer, and four plausible but unequivocally incorrect distractors to reduce guessing (EBEL & FRISBIE, 1986). Even if the items come from a publisher's test bank, how does the instructor knows that the items are high-quality?

Item analysis reveals significant facts about a test, including technical flaws and errors of judgment made by the item writer, multiple interpretations of ambiguous items, and student misconceptions (EBEL & FRISBIE, 1986). In order to improve test validity, instructors must analyze multiple choice items *ex post facto* and use that information to modify or eliminate

poor items for subsequent tests. In time, a large pool of high quality items will allow the instructor to better measure student content knowledge. Calculating both the items' difficulty and discrimination is a simple way to analyze the items.

The proportion of students who got the item correct indicates the difficulty of a multiple choice item (AIRASIAN, 2001). Using GRONLUND (1968) notation for item difficulty (P), the number of students who got the item right (R), and the total (T) number of students who tried a specific item:

$$P_i = \frac{R_i}{T_i}$$

Item difficulty values close to chance levels (25% for a four-option item or 20% for a five-option item) are commonly associated with very hard items, while values closer to unity are commonly associated with easy items (HALADYNA, 1994). An example of an item with a high difficulty index

($P_i \approx 0.80$) from a research-based electricity and magnetism assessment tool (DING, CHABAY, SHERWOOD & BEICHNER, 2006) is:

> *Two small objects each with a net charge of +Q exert a force of magnitude F on each other. We replace one of the objects with another whose net charge is +4Q. What is the magnitude of the force on the +Q charge now?*

An example of an item with a low difficulty index (P_i ˜ 0.20) from the same source is:

> A proton moves with constant velocity $\vec{V}$ to the right through a region where there is a uniform magnetic field of magnitude B that points into the page. There is also an electric field in this region. What is the magnitude of the electric field?

The analysis of item difficulty determines how useful the items are in ranking students by content knowledge. A very easy item that all students can answer does not help the instructor to differentiate between students. The effect of easy items is to add the same amount of points, raising all students' scores. A very hard item that almost no one can answer does not help either. Items of moderate difficulty level contribute most to discriminating among students who have learned varying amounts of subject matter (EBEL & FRISBIE, 1986). AIRASIAN (2001) explains the implication of item difficulty for the standard deviation of a test:

> The difficulty of test items is related to the spread of the scores … When the difficulty of test items is around 50% the resulting test scores will be maximally spread out from low to high. The more pupils' scores differ, the better for making comparisons and distinctions among them (p. 410).

This author also argues that moderate item difficulty is essential for commercial standardized tests but less critical for classroom tests that tend to be criterion-referenced.

If students are known to differ in their performances, then each test item should mirror their tendency to vary (HALADYNA, 1994). Item discrimination is a characteristic of an item that addresses its ability to measure sensitively individual differences by comparing the difference in performance of upper or above average (U) students and low or below average (L) students on a given item (AIRASIAN, 2001; HALADYNA, 1994).

Item discrimination was first described using classical test theory in JOHNSON (1951) and has been reported as a useful and powerful measure by many researchers (ENGELHART, 1965). Using GRONLUND (1968) notation for the item discrimination index (D_i≈0,60) and item difficulty (P), the item discrimination index for a specific item is defined as:

$$D_i = P_{U(i)} - P_{L(i)}$$

An item has positive discrimination when above average students on the test answer it correctly compared with students who performed less than average on the same test. An example of an item with a high discrimination index ($D_i \approx 0.60$) from a research-based electricity and magnetism assessment tool (Ding, Chabay, Sherwood & Beichner, 2006) is:

> In a certain region of space there is a uniform electric field of magnitude E in the +x direction. What is the potential difference V3 – V1 , where location 3 is a distance h vertically below location 1?

An example of an item with a low discrimination index (Di ˜ 0.00) from the same source is:

> Salt water contains n sodium ions (Na^+) per cubic meter and n chloride ions (Cl^-) per cubic meter. A battery is connected to metal rods that dip into a narrow horizontal pipe full of salt water, with the positive end of the battery connected through an ammeter to the right end of the pipe. The cross-sectional area of the pipe is A. The magnitude of the drift velocity of the sodium ions is V_{NA} and the magnitude of the drift velocity of the chloride ions is V_{CL}. Assume that $V_{NA} > V_{CL}$. (+e is the charge of a proton.) What is the correct algebraic expression for the magnitude of the ammeter reading?

Ideally, items should have the largest discrimination index possible, which implies that items are good at ranking students by subject matter knowledge. A discrimination index close to zero but positive suggests that the item is not differentiating student knowledge too well. This is not alarming for classroom tests, but it raises questions about whether the item should be there in the first place if it is not doing its "differentiating" work. Note that in some cases, negative item discriminations can occur. This is a strong signal that the item is flawed, confusing, or that it was keyed incorrectly. On Table 1, Ebel & Frisbie (1986) suggest the following discrimination index cutoff points for norm-based tests with a large sample size.

Table 1

Indexes of discrimination cutoff points for standardized test items

Discrimination index	Item evaluation
1.00-0.40	High discrimination, no need for revision
0.39-0.30	Reasonable discriminating items but possibly subject to improvement
0.29-0.20	Marginal discrimination, usually needing and being subject to improvement
0.19 or less	Poor discrimination, to be rejected or improved by revision

In Table 2, more realistic cut off points for instructor-made multiple choice items are suggested by the author. This table accounts for smaller sample sizes and criterion-referenced tests.

Table 2

Suggested indexes of discrimination cutoff points for instructor-made items

Discrimination index	Item evaluation
1.00-0.30	High discrimination, no need for revision
0.29-0.15	Reasonable discriminating item, revise if possible
0.14-0.00	Marginal discrimination, revisions are recommended
negative	Poor discrimination, to be rejected or revised significantly

A word of caution about the analysis of test parameters is in order, especially when a random sampling of test takers is not viable. Item difficulty is not a constant for a given item. It depends of the characteristics of the students taking the test. As Haladyna (1994) points out:

If the sample contains well instructed, highly trained, or well developed persons, the tests and its items appear very easy, usually above 0.90. If the sample contains uninstructed, untrained, or underdeveloped persons, the test and the items appear very hard … [Item difficulty] is very difficult to estimate accurately unless you are testing a very representative group of test takers (p. 145).

Also, under certain circumstances, item discrimination is underestimated if certain conditions are met, such as if the range of scores is restricted, when instruction is highly effective, or when student effort is high.

Furthermore, researchers have always known that sample size is an important issue to consider in many statistical and test analysis procedures, including item difficulty and discrimination. The smaller the sample size is, the larger the sampling error (Ebel & Frisbie, 1986). These authors describe situations in which a highly discriminating item for a given sample might have low or negative discriminating indexes for another sample of different size. However, even with a small sample size and nonrandom samples, test analysis is still deemed "worthwhile as a means of overall test improvement" (p. 230).

PURPOSE AND RATIONALE

This study applies the concepts of item difficulty and discrimination in the context of assessments used on a general education physical science laboratory by analyzing 72 multiple choice test items. Two research questions guided this study:

1. To what extent the analysis of item difficulty identifies potentially problematic tests items used in lab quizzes?
2. To what extent the analysis of item discrimination identifies potentially problematic test items used in lab quizzes?

As the author of most of the multiple choice test items used for student assessment in the physical science lab and as a trained professional in the basic techniques of science assessment, I hypothesize that the analysis of item difficulty and discrimination will not identify a large number of items as problematic within the context of classroom testing (not necessarily from the perspective of large scale standardized tests).

This type of study is important for a variety of reasons. First, it is designed to improve the quality of the assessment in the general education physical science lab. In addition, it contributes to the physics education literature on assessment. Finally, researchers have noted that studies that examine and analyze item difficulty and discrimination in the context of classroom tests are scarce and "a promising research topic" (Haladyna, 1994, 146).

METHODS

One hundred and seven students, 53 males and 54 females, were enrolled in five sections of the Physical Science Laboratory assigned to the same instructor during the Spring & Fall 2006 semesters. Of these students, about 32% were freshmen, 40% were sophomores, 19% were juniors, and 9% were seniors. These students took the same quizzes at the end of the period, after completing the assigned laboratory of the day. The quizzes consisted of 12 multiple choice items, including conceptual and application items, calculation problems, and graphical analysis. Each item has one correct answer and four distractors. The instructor allowed the students to use their lab manuals as a reference. The author analyzed a total of 72 multiple choice items.

For each quiz, the author found the average and sorted the quizzes into two groups: students who scored above average (U) and students who scored below average (L). Each quiz was analyzed to calculate how many students answered each item correctly for both groups. The author then used the data to calculate each item's index of discrimination, item difficulty indexes for above average and below average scorers, and the overall item difficulty.

For interpretation purposes, the author divided the difficulty index into three categories: easy items (1.00-0.80), moderately difficult items (0.79-0.40), and difficult items (0.39-0.20). The discrimination index can be divided into three categories: high discrimination (1.00-0.30), reasonable discrimination (0.29-0.15), and marginal discrimination (0.14-0). Negative values for this statistic are problematic because below average scorers perform better than above average ones, which is counterintuitive. The items must be carefully studied, modified, or removed from the item pool.

RESULTS

Difficulty index

For above average scorers, about 83% of the items were considered easy. The other 17% of the items were considered moderately difficult. The author identified no difficult items. In contrast, for below average scorers, about 24% of the items were considered easy, 56% of the items were considered moderately difficult, and the remaining 20% of the items were considered difficult (see figure 1). Combining both groups in a weighted average, we obtain 31 easy items, 37 moderately difficult items, and 4 difficult items.

Two examples of very easy items from the data analysis are:

For a group of students, what is the best way to describe the relationship between height and birth month?

a. no apparent relationship

b. proportional linear

c. inversely proportional linear

d. too complex to determine

e. all points fit within a straight line

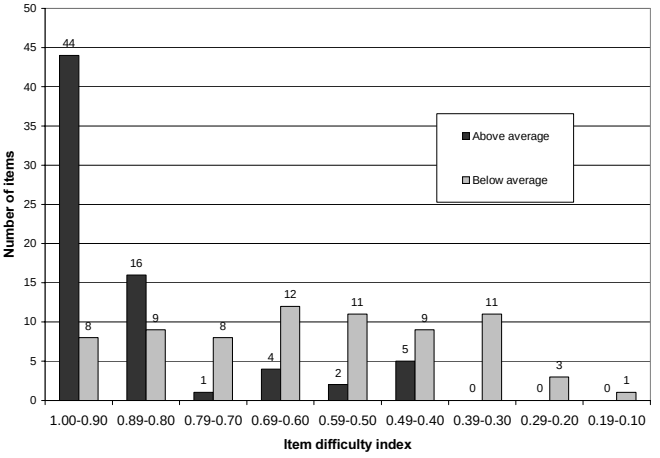


Figure 1. Item difficulty indexes for "above average" and "below average" students

A force that resists motion is known as:

a. friction

b. tension

c. gravity

d. acceleration

e. centripetal force

Two examples of very difficulty items from the data analysis are:

The USS Missouri, an Iowa-class battleship, the mass 4.1 x 10^{10} grams (unloaded). Its volume must be:

a. less than 4.1 x 10^{10} cm^3

b. exactly 4.1 x 10^{10} cm^3

c. more than 4.1 x 10^{10} cm^3

d. cannot be determined without knowing the vessel's density

e. cannot be determined without knowing the density of salt water

In the figure, if the block slides with an acceleration of exactly 1.0 m/s2 on surface A, what would be a possible value for acceleration on surface B?

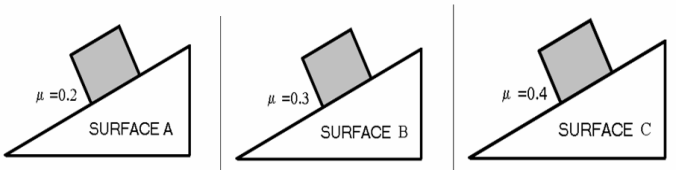
a. 1.5 m/s^2

b. 0.8 m/s^2

c. 1.0 m/s^2

d. 2.0 m/s^2

e. 10.8 m/s^2



Discrimination index

Data suggest that about 43% of the items provide high discrimination between above and below average scorers. About 29% of the items provide reasonable discrimination and the remaining 28% are marginal discriminators. None of the items analyzed have negative discrimination (see figure 2).

From the analysis, two examples of items with a high discrimination index are:

A cinema projector (e.g. the Picwood) uses a lens to focus a frame of film located 0.5 meters from the lens to a screen located 20 meters from the lens. What is the focal length of the lens?

a. 4.88 m

b. 2.05 m

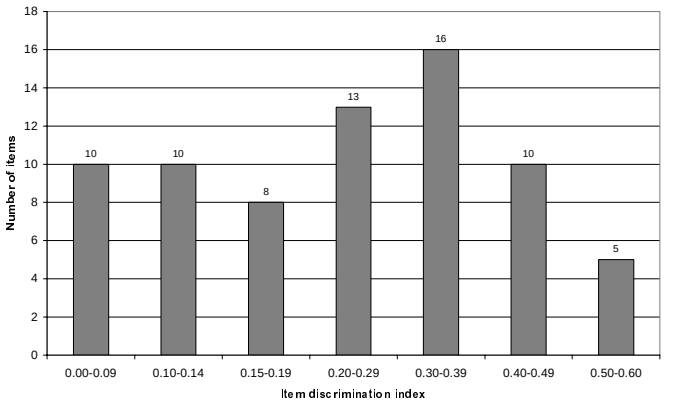


Figure 2. Distribution of item discrimination indexes.

c. 0.488 m

d. 20 m

The collision between cart A (mass = 0.20 kg) and cart B (mass = 0.27 kg) is represented in the following graph. Cart B was at rest before the collision, just like in our laboratory today. What is the total momentum after the collision?

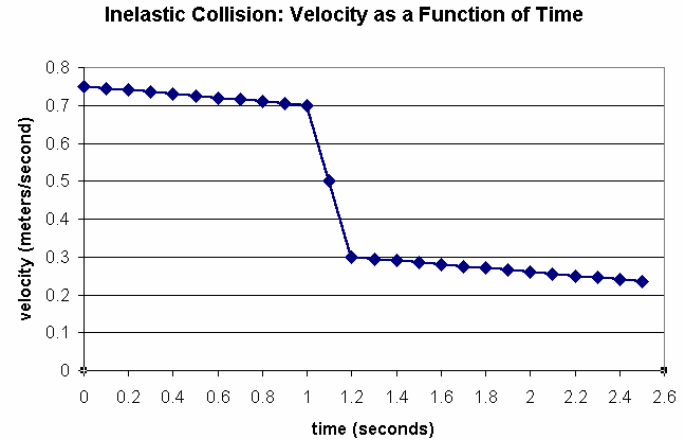
a. 0.7 m/s

b. 0.049 kg m/s

c. 0.049 J

d. 0.14 J

e. 0.14 kg m/s



Two examples of items with low discrimination index are:

Which of the following is considered a wavelength of the visible portion of the electromagnetic spectrum?

a. 800 nm

b. 950 nm

c. 250 nm

d. 300 nm

e. 550 nm

The spectrum of an incandescent light bulb looks very much like a rainbow. What type of spectra is this?

a. emission

b. absorption

c. striped

d. continuous

e. refraction

Difficulty and Discrimination

Of the 72 items analyzed, 16 of them can be classified as both easy and marginally discriminant, that is, both above and below average scorers found them too easy. Instructors must carefully examine items like these because they do not contribute to ranking students based on their knowledge. Also, 3 items with marginal discrimination were considered moderately difficult and 1 item was considered difficult for all scorers. In this case, the difficult item is not helping in ranking students either.

Out of 22 items with reasonable discrimination, 12 can be classified as easy, 8 as moderately difficult, and 2 as difficult. It would not be a bad idea to examine some of the easy items in this category to search for ways to make them at least moderately difficult without affecting their discrimination index.

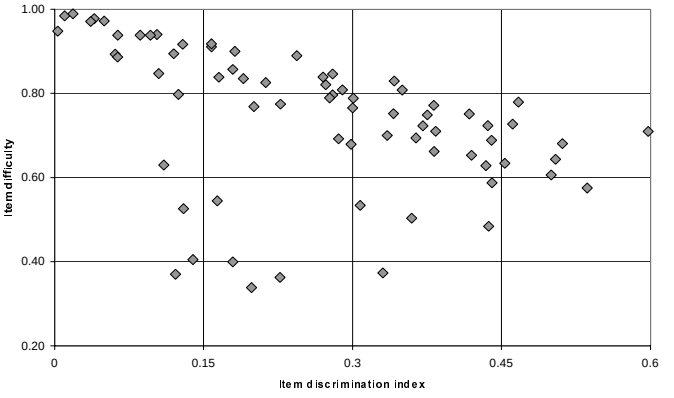


Figure 3. Item correspondence between item difficulty and item discrimination indexes.

The rest of the items, about 30, fall into the category of high discrimination. Of those items, 27 are classified as moderately difficult. According to the literature, these items are maximizing student ranking by content knowledge and should not be modified (see figure 3). An example of these type of items (Pi = 0.70; Di = 0.33) is:

If the momentum before and after a collision between two carts is the same, the collision can be classified as:

a. it could be elastic or inelastic

b. inelastic only

c. elastic only

d. cannot be answered without knowing the mass of the carts

e. cannot be answered without knowing the speed of the carts

Another example of an item with good item difficulty and high item discrimination (Pi = 0.64; Di = 0.50) is:

What is the slope of this line?

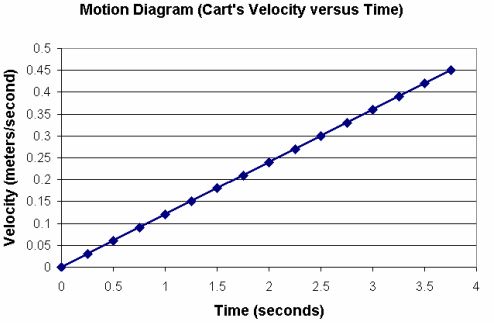
a. 0.05 m/s^2

b. 0.12 m/s^2

c. 0.55 m/s^2

d. 8.33 m/s^2

e. 9.8 m/s^2



DISCUSSION

The author's analysis of item difficulty and discrimination has allowed an objective impression of item quality from the perspective of how the students reacted to them, which is not always the same as the instructor intended the items to work. The results reveal that despite the author's efforts to write good items, about 22% of them might not be considered highly effective at their ultimate purpose: accurately measure students' understanding of physical science. Since there are many more easy items compared with difficult one, the scores are biased toward the higher end of the scale. At least some of these flagged items should be evaluated and modified to prevent "giving away" points.

A possible explanation for this many items being flagged is because the quiz is open-book. The idea of having open-book quizzes, which is not always seen as prudent in the education literature (Clift and Imrie, 1981; Crooks, 1988), came from a majority of the instructors who teach other sections of the physical science lab regularly. They argued that it was not a realistic expectation for students to listen to the pre-lab, complete the laboratory experience, listen to the post-lab summary, and "absorb" enough material to succeed on a closed-book quiz in less than two hours.

Having about 22% of the items flagged is not problematic in the case of classroom tests because they are not norm-based and will not be graded on a curve. On the other hand, about 78% of the items have appropriate difficulty and/or discrimination. It is these items that are carrying most of the weight of ranking students by grade in the lab sections studied.

IMPLICATIONS

Some suggestions from the literature and from the author's personal experience analyzing the test items include:

1. High difficulty items must be carefully examined. A high difficulty item is not necessarily an invalid one. If an unequivocally incorrect distractor is selected by most students, it could mean that the particular concept was not taught properly, or that it was not understood by students. Test scores should not be automatically raised (or an item eliminated) just because many students got an item wrong (Airasian, 2001).
2. Low difficulty items must be carefully examined, but for a different reason. A low difficulty item is not necessarily an invalid one. If the teacher want to be sure that all students know the very essential concepts, easy items testing those concepts are acceptable.
3. Closed-book tests are better at testing what the student learned and remembered from the laboratory. If an open-book test will be used, make sure that none of the test items are directly answered in the laboratory manual. Instead, write comprehension, interpretation or analysis questions.
4. After analyzing discrimination and difficulty, the analysis of individual distractors is the next logical step. Some questions that might be asked are: 1) Why do students choose a particular incorrect distractor? 2) Why are some distractors never chosen by students? Since test performance is affected, among other things, by the quality of the distractors, Haladyna (1994) recommends a thorough analysis of them for sound item and test development. For example, distractors that are seldom or never chosen and extremely implausible distractors should be replaced.
5. The perceived difficulty of a test is related to how well prepared students are for it. For the same item, well prepared students see it as easier and less prepared students see it as difficult. In fact, the average difficulty for all 72 test items for above average scorers (0.86) compared with below average scorers (0.60) is statistically different (paired $t = 14.85$, $p < 0001$), just as if they were answering completely different tests.

CONCLUSIONS

This paper applies item difficulty and discrimination to analyze the quality of the multiple choice test items used to grade students enrolled on the "Introduction to Physical Science Laboratory" course. The analysis identified several potentially problematic items with high difficulty index (very easy items) and/or low discrimination index (students performed about the same regardless of overall test score). The flagged items will be examined and modified, if possible. Another option is to eliminate them since they do not contribute to the overall evaluation goal of the test.

The analysis of multiple choice items using difficulty and discrimination can be a time consuming task, but it is an important one. If we want to assign a grade that correlates to the students' mastery of the subject matter, we must not trust exclusively the instructor's ability to write good multiple choice items. In some cases, students read an item from a completely different perspective than the instructor. An instructor will never really know how well test items will work until they have been administered to students and analyzed after the fact using some of the techniques discussed in this paper. Unlike many quantities in physics, measuring student knowledge is fraught with confounding variables associated with the student, the instructor, the test itself, and the testing environment. By choosing multiple choice items with optimal difficulty and discrimination, physical science instructors can develop the most effective and valid assessments possible.

BIBLIOGRAPHY

AIRASIAN, P. W. Classroom assessment: Concepts and applications. Boston, MA: McGraw Hill, 2001.

EBEL, R. L. & FRISBIE, D. A. Essentials of educational measurement, 4th ed. Englewood Cliffs, NJ: Prentice Hall, 1986.

CLIFT, J. C. & IMRIE, B. W. Assessing students, appraising teaching. New York: Wiley, 1981.

CROCKER, L. M. & ALGINA, J. Introduction to classical and modern test theory. Fort, 1986.

WORTH, TX: Harcourt Brace Jovanovich.

CROOKS, T. J. The impact of classroom evaluation practices on students. Review of Educational Research, 58 (4), 438-481, 1988.

DANCY, M. H. & BEICHNER, R. Impact of animation on assessment of conceptual understanding in physics. Physical Review Special Topics - Physics Education Research 2, 010104, 2006.

DING, L.; CHABAY, R.; SHERWOOD, B. & BEICHNER, R. Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. Physical Review Special Topics - Physics Education Research, 2, 010105, 1-7, 2006.

ENGELHART, M. D. A comparison of several item discrimination indexes. Journal of Educational Measurement, 2 (1), 69-76, 1965.

ENGER, S. K. & YAGER, R. E. Assessing student understanding in science: A standards-based K-12 handbook. Thousand Oaks, CA: Corwin Press, 2001.

GRONLUND, N. E. Readings in measurement and education. London, UK: Collier-Macmillan Ltd., 1968.

HALADYNA, T. M. Developing and validating multiple-choice test items. Hillsdale, NJ: Lawrence Erlbaum Associates. 1994.

HAZEL, E., LOGAN, P. & GALLAGHER, P. Equitable assessment of students in physics: Importance of gender and language background. International Journal of Science Education, 19 (4), 381-392, 1997.

HEDGES, W. D. Testing and evaluation in sciences in the secondary school. Belmont, CA: Wadsworth Publishing, 1966.

HOGAN. T. P. Educational assessment: a practical introduction. Hoboken, NJ: Wiley, 2007.

JOHNSON, A. P. Notes on a suggested index of validity: The U-L index. Journal of Educational Psychology, 42 (8), 499-504, 1951.

JOHNSTONE, A. H. & AMBUSADAI, A. Fixed response: What are we testing? Journal of Science Education/REC, 2 (1), 30-31, 2001.

MINTZES, J. J., WANDERSEE, J. H. & NOVAK, J. D. Assessing science understanding: A human constructivist view. San Diego, CA: Academic Press, 2000.

NITKO, A. J. Educational assessment of students. Englewood Cliffs, NJ: Prentice Hall, 1996.

O'BRIEN-PRIDE, T., VOKOS, S. & MCDERMOTT, L. C. The challenge of matching learning assessments to teaching goals: An example from the work-energy and impulse-momentum theorems. American Journal of Physics, 66 (2), 147-157, 1998.

RACE, P. Why do we need to "repair" our assessment procedures? A discussion paper. Journal of Science Education/REC, 2 (4), 73-76, 2003.

SHULMAN, L. S. Those who understand: Knowledge growth in teaching. Educational Researcher, 15 (2), 4-14, 1986.

SLATER, T. F.; RYAN, J. M. & SAMSON, S. L. Impact and dynamics of portfolio assessment and traditional assessment in a college physics course. Journal of Research on Science Teaching, 34 (3), 255-271, 1997.

THORNDIKE, R. L.; ANGOFF, W. H. & LINDQUIST, E. F. Educational measurement, Washington, D.C.: American Council on Education, 1971.

THORNTON, R. K. & SOKOLOFF, D. R. Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula. American Journal of Physics, 66 (4), 338-352, 1998.

Received: 11.05.2007 / Approved: 20.04.2008