Effects of item writing rules on the reliability of instruments to measure the mathematical knowledge of teachers

Jim Gleason, Ph.D. †
Ashley Alley ‡
Stephanie Baker §

Abstract

This project analyzes the importance of item-analysis on the ability to gather useful item information. As an item's information is difficult to gather prior to administering a test, a method of accurately predicting how much information an item will bring is useful in test construction. While the scope of this project is narrow focusing on merely thirty items, the results are statistically significant. Using multiple regression, this study focused on showing that item violations significantly effects item information ($p = 0.0227$). Furthermore, the combination of difficulty and item violations significantly affect item information ($p = 0.014$). Thus, item violations and difficulty of an item are important in determining an item's overall validity and information. Further studies are needed in order to expand the scope of this project, and more items should be analyzed in a more in-depth project.

Introduction

Creating instruments to measure individuals' abilities in certain cognitive areas is an important component of our educational system. While most educators prefer free-response questions to true-false and multiple-choice, there are times where for economic reasons (time, money, etc.) the fixed-response format is necessary. Although item-writing is still mostly a creative art, as stated by Ebel (1951), using both a teacher's creative energies as well as his or her content knowledge, a standard set of guidelines should be followed to stabilize and standardize the reliability and validity of test items (Haladyna, Downing, & Rodriguez, 2002).

The concern of some individuals regarding the ability to test higher-order thinking skills with multiple-choice test options is causing many teacher-made and norm-referenced tests to move towards essay and free-response format. However, with much work and by following item-writing rules based upon research and experience, one can test the same types of skills and knowledge as with essay or free-response items (Klein, 1998). Much of the controversy stems from the need to simplify items so that they remain stable and valid. However, if authors follow appropriate item-writing guidelines, then items may retain their higher cognitive order without losing any of their information. Because multiple-choice formats result in more economical and faster scoring, these formats are preferable, especially on norm-referenced exams.

The ultimate goal when creating a measurement instrument is to provide as much information about a subject's ability from each item, thereby contributing to the entire instrument's reliability in distinguishing between the abilities of different subjects. While classical test theory does not provide ways in which to measure the amount that each item contributes to the overall instrument reliability, item response theory is based upon finding the amount of information that each item measures and then adding them together to find the instrument's reliability (Hambleton, Swaminathan, & Rogers, 1991). Therefore, the use of item response theory in the analysis of a measurement instrument is critical to conduct research regarding item-writing skills.

Due to the theoretical construct of most item response theory models, the difficulty of an item directly affects the amount of information the item provides when it is very easy or very hard and has little impact when the item difficulty is in the middle of the range. This likely accounts for the research that item difficulty has been found to be the least important variable when determining validity of an item (Haladyna & Downing, 1989). Additionally, since the difficulty of an item is best determined after an item's use in an instrument with a large population, the same time that full item information is achieved, its use is limited in the item creation process.

In an effort to improve the item writing process, many taxonomies and guides exist regarding the development of multiple choice test items according to a standard set of item-writing rules; however, research regarding item-writing procedures is still very young with few results (Haladyna, Downing, & Rodriguez, 2002). Since "item difficulty is very sensitive to changes in how questions are asked" (Millman & Greene, 1993, p.348), item-writing rules are an integral part of determining the validity and item information of a test item. However, in comparison to other fields which affect validity and reliability, research regarding item-writing rules is scant (Haladyna & Downing, 1989a).

Many sets of item-writing rules have been developed specifically for this use of test development (Haladyna & Downing, 1989a). In a related study, Haladyna and Downing (1989b) validated the guidelines which they established through a two-pronged method. They first conducted a literature review to search for articles containing empirical research regarding item-writing rules. They then consulted 20 of the most popular texts regarding classroom assessment in search of author consensus. In a later study, Haladyna, Downing, and Rodriguez (2002) call for more research on item-writing guidelines and validity of such guidelines. In response to this call, the Content Knowledge for Teaching – Mathematics instrument is examined regarding its adherence to these item-writing guidelines to determine the effect of the violations of the item-writing rules upon the low reliability of the instrument when used in evaluating universities' elementary education programs.

Method

Apparatus

In the mid 1980's, Lee Shulman communicated a construct of knowledge for teaching into the three major components of "(a) subject matter

content knowledge, (b) pedagogical content knowledge, and (c) curricular knowledge" (1986, p. 9). In the specific subject of mathematics, several groups have designed instruments to measure this knowledge for teaching during the past decade (Brown, McGatha, & Karp, 2006; Hill, Schilling, & Ball, 2004; Ferrini-Mundy, McCrory, & Senk, 2006). For this study, we focus on the Content Knowledge for Teaching – Mathematics (CKT-M) instrument developed by the Learning Mathematics for Teaching (LMT) project at the University of Michigan (Hill, Schilling, & Ball, 2004).

As part of the Mathematical Education of Elementary Teachers (ME.ET) project at Michigan State University, forms using items from the CKT-M instrument were designed for and given to over 750 pre-service teachers in Michigan, South Carolina, and New York City (McCrory et. al., 2008). Using these responses, we established item difficulty and item information parameters for each of the items on the forms.

Since items in which almost all subjects either answer correctly or incorrectly produce little information as to the subjects' abilities, the item difficulty parameter used for this project is based upon how well an item measures the middle group of students. Therefore, for each item that had only one part, the item difficulty parameter is the absolute value of the difference between the percent correct and 0.50. For testlet items (items with a single stem and multiple questions), the item difficulty parameter is the average of the difficulty parameters for each part.

Rather than using a single reliability coefficient to measure an instrument's reliability, item response theory creates a test information curve showing reliability at various locations along the ability spectrum as measured by the instrument. This test information curve is the sum of the item information curves constructed using the parameters for each of the items. For the current study, a three-parameter model for non-testlet items and a graded model for testlets generated an item information curve for each of the items analyzed. The item information parameter used for the study is the average value of the item information function between three standard deviations below and above the mean of the subject scores.

Procedure

After all the preliminary information was gathered about the items, we needed a way to predict when a newly made item would be reliable. As difficulty could only be measured by how the students performed on the test, it was too expensive to use item difficulty to predict reliability. A method to test the items prior to using them was important to develop. We chose multiple regression due to its ability to check several independent variables against a dependent criterion. The dependent criteria, an item's information, remained the same throughout the research. This statistical procedure was chosen as a manner in which to test ideas about how an item could be made to be reliable and because it allows a person to predict an outcome, in this case when an item will be reliable.

The next task was to determine which criteria would affect the reliability of the study. This portion of the study began with an idea that the type of math (e.g. geometry, number theory, basic arithmetic, etc) would affect the information of an item; however, this approach did not have the expected results. While the items could successfully be categorized into a few mathematical groupings, these collections were not independent of one another, and thus any statistical data gathered was unreliable and inaccurate when making predictions.

While trying to decide which criterion would affect the overall item information, we decided to consider how the items were written. Multiple choice and true false questions have recommended rules when writing each item. Using results of Haladyna and Downing (1989b), each item was marked by which rules it violated (for a list of the item-writing rules used, see Table 1). Two of the authors independently went through the sets of items to determine which rules were violated. After each individual had been through all of the items, the results were compared and debated until appropriate score was agreed upon for each item.

The frequency in which the item-writing rules appeared in classroom assessment textbooks was used for the weight that the particular violation received (See Appendices A and B). If a rule also had research support, its weight was automatically elevated to a 1. Thus, the criteria, "All answer options should be plausible" received a .70 while "order of answer options should be logical or vary" received a 1 due to its research support. There was one exception to this rule; one criterion which did not have research support was also elevated to a 1. This criterion was "answer options should include only one correct answer". This particular element is basic and fundamental to an item and so the weight was raised to a 1 even without research support. Furthermore, rule 29, "Vague frequency terms (e.g. often, usually) should not be used", and rule 9, "Specific determiners (e.g. always, never) should not be used", have similar reactions among test-takers and so were both given the higher weight of 0.5.

Each item in the CKT-M instrument was then analyzed according to the number of item-writing rules violated and given a score based upon the weights of the rules violated. Furthermore, the frequency of the violations was taken into account by adding up how many times a particular rule was violated within the same item. For example, if an item has more than one implausible answer, the item would be marked for each answer that violates this rule. For further examples on how the items were scored, see the two items below.

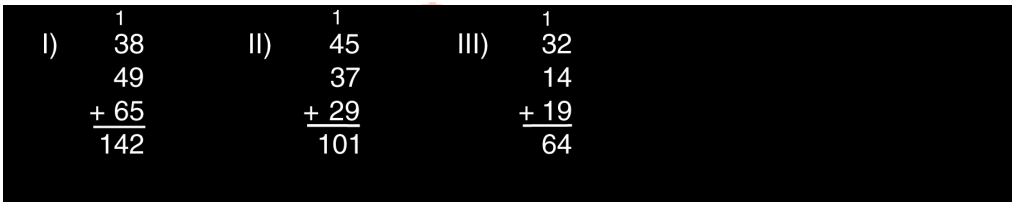
Ms. Harris was working with her class on divisibility rules. She told her class that a number is divisible by 4 if and only if the last two digits of the number are divisible by 4. One of her students asked her why the rule for 4 worked. She asked the other students if they could come up with a reason, and several possible reasons were proposed. Which of the following statements comes closest to explaining the reason for the divisibility rule for 4? (Mark ONE answer.)

Four is an even number, and odd numbers are not divisible by even numbers.

The number 100 is divisible by 4 (and also 1000, 10,000, etc.).Every other even number is divisible by 4, for example, 24 and 28 but not 26. It only works when the sum of the last two digits is an even number.

This item only violates one rule. Rule 29 which states, "Vague frequency terms (e.g. often, usually) should not be used" is violated because the item asks "which of the following statements comes *closest* to explaining the reason". "Comes closest" is vague and could potentially lead to more than one answer which could be deemed correct. It is a phrase that leads to subjectivity which lowers the validity of the instrument. This item would receive 0.50.

Mrs. Jackson is getting ready for the state assessment, and is planning mini-lessons for students focused on particular difficulties that they are having with adding columns of numbers. To target her instruction more effectively, she wants to work with groups of students who are making the same kind of error, so she looks at a recent quiz to see what they tend to do. She sees the following three student mistakes:



Which have the same kind of error? (Mark ONE answer.)
I and II
I and III
II and III
I, II, and III

This item violates Rules 1 and 37. Rule 1 states, "'All of the Above' should not be an answer option". This rule has research to support it so the item would be scored as a 1 from this rule. Rule 37 says, "Complex item formats ('a and b, but not c') should not be used". This item-writing rule is also research supported and thus also has a 1 for its point value. Therefore, the total point value of this item is 2. Items used in this research had scores ranging from 0 to 3.85 so this item ranks high in item violations. See Table 2 for a listing of the items, their item information score, difficulty, and item-writing violation score.

After rating each item and determining its overall score, adding in all violations made by a particular item, multiple regression was run in order to determine this criterion's impact on the item's information. The item difficulty parameter was also added in as a factor because an item's difficulty is a variable effecting item information as well. After running multiple regression using difficulty and item writing violations as the two independent variables with item information as the dependent variable, it was found that the item writing violations does significantly have an influence on the information of an item.

Results

Using an alpha value of .05, the model summary and the ANOVA summary indicate the overall model of the two independent variables significantly predict item information, $R^2 = 0.2710$, $R^2adj = 0.2170$, $F(2, 29) = 5.0190$, $p = 0.0140$. However, review of the beta weights specify that only one variable, item writing violation score $\beta = -0.3485$, $t(29) = -2.4161$, $p = 0.0227$ significantly contributed to the model. The β being negative corresponds to the increase in item-writing violation score corresponding to a decrease in item information, with the low magnitude stating that the influence is not very strong.

Even though item-writing violation was the only criterion that was significant in the model, it is worth noting that difficulty was barely outside the range of significance. Difficulty $\beta = -2.7179$, $t(29) = -1.9976$, $p = 0.0559$ was very close to aiding to the model using an alpha of 0.05 with the magnitude of β being much larger and therefore likely contributing more to the information of an item. This matches the hypothesis that the distance of an item's difficulty from the average corresponds with the item containing less information. See Table 3 for the statistical output.

Discussion

Significant results were found showing that item-analysis plays an important role in determining an item's information; however further research with more item analysis is needed in order to gain a better understanding of the effect of item analysis on information gathered from a test. Item-analysis is clearly important as it is statistically significant even with the limited number of items that were available to analyze. It should be noted that when looking for item violations, an item cannot merely be looked at superficially. One must closely examine each item and each response choice in order to determine every item violation that is made. Some item-violations are subtle and cannot be found on the first read through. Each item analyzed in this project was looked at multiple times which allowed us to get a more in-depth look at each item. Furthermore, supplementary research should be done using more items to gain additional reliability. Expanding the scope of the project to include multiple tests so the items are less similar to each other would allow one to broaden the use of the results to further applications.

† *Jim Gleason, Ph.D.,* The University of Alabama, Alabama, USA
‡ *Ashley Alley,* The University of Alabama, Alabama, USA
§ *Stephanie Baker,* The University of Alabama, Alabama, USA

References

Brown, E. Todd, McGatha, M., Karp, K. (2006) Assessing teacher knowledge: Diagnostic assessments for elementary and middle school teachers, *New England Journal of Mathematics*, 37-50, May.

Ebel, R. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407-424.

Ferrini-Mundy, J., McCrory, R., & Senk, S. (2006). *Knowledge of algebra teaching: Framework, item development, and pilot results*. Research symposium at the research pre-session of NCTM annual meeting. St. Louis, MO.

Haladyna, T.M., & Downing, S.M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50.

Haladyna, T.M., & Downing, S.M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 51-78.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal*, 105 (1), 11-30.

Klein, S.P. (1998). Standards for teacher tests. *Journal of Personnel Evaluation in Education, 12*(2), 123-138.

McCrory, R., Cole, Y., Lovin, L.A., Moss, M., Smith, S., Costner, B., Pullano, F., Beckmann, S. (January 2008). *Mathematics Courses for Elementary Teachers: An Overview of Current Research Projects*. Presentation at annual meeting of Association of Mathematics Teacher Educators, Tulsa, OK.

Millman, J., & Greene, J. (1993). The specifications and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational measurement*, (3rd ed.). Phoenix, AZ: American Council on Education.

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