The Validity of Item Bias Techniques with Math Word Problems

Gail Ironson, Susan Homan, and Ruth Willis
University of South Florida

Barbara Signer
University of Massachusetts

Item bias research has compared methods empirically using both computer simulation with known amounts of bias and real data with unknown amounts of bias. This study extends previous research by "planting" biased items in the realistic context of math word problems. "Biased" items are those in which the reading level is too high for a group of students so that the items are unable to assess the students' math knowledge. Of the three methods assessed (Angoff's transformed difficulty, Camilli's full chi-square, and Linn and Harnisch's item response theory, IRT, approach), only the IRT approach performed well. Removing the biased items had a minor effect on the validity for the minority group.

Test bias is a serious issue for educators as well as the general public. Statistical methods for detecting bias in test items have been developed for both practical and theoretical reasons; that is, in the absence of an external criterion they may provide the only means available for test developers to screen for potential bias (Rudner, Getson, & Knight, 1980a), but they also allow an evaluation of the internal structure and construct validity of the test (Shepard, 1981).

Item bias research has been focused on the development of specific indices (e.g., Lord, 1977; Scheuneman, 1979) and on their experimental application to particular tests. The conceptual and mathematical properties of the procedures have been appraised (Hunter, 1975; Peterson, 1977), and empirical studies have been done comparing methods using both computer simulation (Rudner, Getson, & Knight, 1980b) and real data (Ironson & Subkoviak, 1979; Shepard, Camilli, & Averill, 1981). Thus far, empirical validation of the bias methods has not been the focus of attention, though several studies have contributed relevant evidence.

Ironson and her colleagues have been following a program of research in which tests are constructed so that the potential source of bias is known (Ironson & Craig, 1982; Subkoviak, Mack, Ironson, & Craig, 1984). Bias techniques can then be evaluated for their sensitivity to these sources of bias with actual populations of test takers.

In the present study, a test was constructed in which the readability of math word problems was manipulated. The validity of the statistical detection techniques would be supported if math problems with difficult reading levels were found to be biased. Another question of interest was whether the test would become more valid for a minority group when the biased items were deleted.

Method

Instrument and Subjects

The test, a mathematical word problem test, consisted of three types of items: (1) 10 items at second grade reading and second grade math level, (2) 10

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 4, Fall 1984, pp. 391–396

0146-6216/84/040391-06$1.55

at second grade reading and fourth grade math level, and (3) 6 items at fourth grade reading and fourth grade math level.

A sample of 1,064 students from second and fourth grades were tested on this instrument. In addition, scores from the reading and math tests of the Comprehensive Tests of Basic Skills (CTBS; 1973) were obtained as an independent assessment of the students' reading and math levels. Subsets of students were then chosen to represent majority and minority groups. The "majority" group (n = 916) was students whose reading and math levels were above fourth grade. The "minority" group (n = 148) was those students who were at fourth grade or above on math level but below fourth grade on reading level. The items that were presumed to be biased for the minority group were the six items at fourth grade reading and fourth grade math levels. These questions were, therefore, at the appropriate math level but at an inappropriately high reading level. It was assumed that the minority group had the requisite math skills to answer these items but that they might not have been able to access that information because the verbal level of the items was too high. Hereafter these items are referred to as the "biased" items.

Evidence for the validity of the math word test is provided by a correlation of .66 with the CTBS math assessment test total scores.

Readability Level

The readability level of the items was determined using the Homan Readability Formula (Homan, 1980). The Homan Readability Formula is a new formula that was designed to examine the readability level of a unit as small as a sentence or test item. The purpose of the formula is to give test item developers a method for determining if an individual test item was written at the grade level for which it was intended.

The Homan Readability Formula utilizes three variables—vocabulary load, sentence complexity, and syntactic density. The variables are used to determine whether an individual sentence or item is written at the intended readability (grade) level.

The first variable, vocabulary load, was determined by using the number of unfamiliar words as delineated in the *Living Word Vocabulary* (Dale & O'Rourke, 1981). This corpus, first completed in 1979 but continually updated, has over 48,000 words and meanings. For each word, one or more definitions are listed, along with the percentage of students at a particular grade level who recognized (were familiar with) that definition for a given word. A word has to be familiar to at least 60% of the students in the sample at the required grade level to be considered familiar. If a word was not listed, it was considered an unfamiliar word. Previously, the only such available lists had words listed by frequency of appearance at various grade levels; word meaning was not included. The added dimension of word meaning greatly enhanced the value of the vocabulary load variable, already the most important of the three variables.

The second variable, sentence complexity, was determined by simply counting the number of words in the sentence. It has long been accepted that, for the most part, the longer the sentence, the more complex the sentence.

The third variable is sentence density. Sentence density is very difficult to measure objectively. Sentence density attempts to measure the difficulty of a sentence in terms of syntactic or grammatical structure. Although sentence length is often considered a measure of complexity (the longer the sentence, the more complex the thought), this is not always true. On occasion, the thought or content of a long sentence is less complex than a short sentence.

The *T*-unit, or "thought" unit developed by Hunt (1970), is a measure quantifying the number of words used to express a thought. It attempts to differentiate between long syntactically complex sentences and long syntactically easy sentences. For example, the sentence "In nature, light makes color, and in art, color makes light" is an 11-word complex sentence. An easy 11-word sentence would be, "The little duck went swimming and the funny dog went swimming." *T*-units can be used to measure the syntactic density of both sentences. To express a thought, the first sentence requires 11 words and therefore has 11 words per *T*-unit. The second sentence expresses two thoughts using 11 words and therefore has 5½ words per T-unit. The simpler sentence has more T-units and fewer words per T-unit. The complex sentence has fewer T-units and more words per T-unit. Thus, the more complex sentence receives a higher T-unit score. This score was used as a measure of sentence density.

These three variables—vocabulary load, sentence length, and sentence density—were combined into a ranking for each sentence, called a sentence score. The fourth grade sentence scores ranged from 9 to 60. The second grade sentence scores ranged from 6 to 32. Based on previous research (Homan, 1980), these sentences (test items) had scores within the acceptable range.

Mathematics Level

All questions were taken from either a diagnostic mathematics test or a state-approved arithmetic testbook. In both cases the appropriate mathematics grade level, either second or fourth, was indicated in the test or book. Additional checks were made to insure a variety of questions.

Second grade level. The second grade level questions were selected from the following sources: (1) the Pinellas County Mathematics Test Grade 2 (Uprichard, 1979–80), and (2) *Elementary School Mathematics Book 2* (1971). The test was chosen because it is one of the few diagnostic mathematics tests for the second grade, in which questions do not have to be given orally to the students. The questions from the test were taken from the applications section. The text was chosen because it commonly appears on approved testbook adoption lists. On close examination of this text by the Homan Readability Formula, the reading level of many of the questions was not beyond the second grade level. This was not the case in the other texts examined. This is not to say that other texts do not meet this criterion. However, the chosen text was the most appropriate of those available. Questions from the chosen text were taken from different sections of the book to insure variety of question format.

All second grade level questions involved the application of a single operation, either addition or subtraction. However, the phrasing of the questions differed.

Fourth grade level. The fourth grade level questions were selected from the following sources: (1) the Pinellas County Mathematics Test Grade 4 (Uprichard, 1979–80), (2) the Stanford Diagnostic Mathematics Test, Brown Level (SDMT; 1976), and (3) The Science Research Associates (SRA) Elementary Mathematics Program, 4th Grade Level (1974).

The Pinellas County Mathematics Test and the SDMT provided the second grade reading level questions. The SRA test was the source for the fourth grade reading level questions. This assignment was made after applying the Homan Readability Formula. These tests and text were considered because of their acceptance by school districts.

The fourth grade level questions met one of the two following criteria: (1) applying two of the four basic operations on whole numbers, or (2) solving by a proportion. Again, consideration was given to vary the phrasing of the questions.

Item Bias Indices

Three item bias indices that had shown promise from previous studies were used in this study. The first, described in Angoff and Ford (1973), uses as a measure of item bias the distance of an item's delta values from the major axis of an ellipse composed of the deltas of all items. The second procedure, described in Ironson (1982), called the chi-square or Camilli chi-square, matches minority and majority groups on total scores. Then chi-square is calculated at each level, a sign indicating the direction of bias is attached, and the chi-square is summed over levels. The third procedure is a latent trait or item response theory (IRT) procedure modified for use with a small minority sample (Linn & Harnisch, 1981). Basically, all cases in the sample are used to estimate each item's parameters. Then based on the ability (or θs) of subgroup members, the observed proportion correct is compared to the model. Differences may then be standardized.

Results

The test was considerably more difficult for the minority group on all three types of items. The average proportion correct (p) for the three types of items, was as follows:

Fourth grade reading, fourth grade math: $p = .26$ majority, $p = .15$ minority;
Second grade reading, fourth grade math: $p = .44$ majority, $p = .25$ minority;
Second grade reading, second grade math: $p = .78$ majority, $p = .58$ minority.

As is well-known in bias research, direct examination of p values is an incorrect procedure since there is no control for ability differences. However, note that the biased items (fourth grade reading, fourth grade math) were more difficult than the other items. They were, in fact, more difficult than the second grade reading, fourth grade math items (this would not be the case if the test were a pure math test). It is also of interest to note that the biased items were not relatively more difficult for the minority group as compared with the other items.

The major bias analysis involved calculating the rank order correlation between each signed bias index and a 0/1 unbiased/biased classification of items (1 = fourth grade reading, fourth grade math items; 0 = all other items). In view of the above p values, it was not surprising to find a small nonsignificant correlation ($r = -.23$, $n = 26$, *n.s.*, $Pr[-.57 \leq p \leq .17] = .95$) between Angoff's item bias measure and the unbiased/biased classification of items. The absolute values of the individual item distance measures were all less than 1.0, suggesting that none of the items appeared to be biased. The correlation for the full chi-square was equally disappointing ($r = -.23$, $n = 26$, *n.s.*, $Pr[-.57 \leq p \leq .17] = .95$), and no individual items would have been identified as biased. Given these results, Linn and Harnisch's standardized difference scores for the majority correlated remarkably well with the bias classification ($r = .63$, $n = 26$, $p < .01$, $Pr[.32 \leq p \leq .82] = .95$). In addition, five of the six biased items had bias indices among the highest seven according to Linn and Harnisch procedure.

A small but noticeable effect of the biased items on the validity of the test was found. The validity was measured by the correlation of the math word problem test with the CTBS math assessment. The correlation of the total test score for the entire group was .661, and with the biased items removed it was .656. This effect was negligible since the validity of a shorter test would be expected to be lower. However, removing the biased items slightly increased the validity for the minority group (from .250 to .283) and slightly decreased the validity for the majority group (from .555 to .536), probably due to the shorter test length. Thus, the biased items did have a minor effect on the validity for the minority group.

Discussion

One reason that the IRT approach may have worked as well as it did is that the higher level reading items may have resulted in some multidimensionality. In fact, bias is sometimes thought of as a kind of multidimensionality involving measurement of a primary dimension and a second confounding dimension. Group differences on the second dimension (reading) are relatively larger or smaller than those on the primary (math) dimension. Some support for this was found in the present data where the mean difference between groups for reading scores on the CTBS (420.79 − 314.39) was greater than the mean difference for math scores on the CTBS (393.71 − 349.15).

In the area of planted bias then, this study and the Subkoviak et al. (1984) study suggest that the IRT procedure is, indeed, the most powerful of the methods studied for detecting this type of bias. Such a general conclusion must be taken as tentative, however, since more studies need to be done, and one study (Ironson & Craig, 1982) found strong support for the Angoff and the one-parameter procedure as well.

Although it may be argued that the estimates of θ which are used to develop the IRT curve are themselves biased, using a redefined estimate of ability independent of the items (such as CTBS math scores) was not pursued for the following reasons:

1. A primary concern in doing item bias studies is that biased items may be missed. However, this study has already found that the IRT approach did quite well at identifying the biased items. Furthermore, if Camilli's and Angoff's procedures performed better with the redefined ability measure (CTBS scores), they still did not do well in practice; so what knowledge does this add?
2. By using CTBS scores to identify items as really biased (i.e., another criterion measure), there would be four definitions of bias—three statistical and the one currently used in this paper (by construction, in which the reading level is too high to access the math ability). Defining bias by statistical procedures has been done quite well many times, and the purpose of this paper was to define bias by construction.
3. This study was directed toward investigating the statistical procedures as they are used in practice, not in some setting with information that researchers almost never have. It is unrealistic to expect that researchers would have an unbiased estimate of what they are trying to measure.

It is hoped that the results of this study go one step further in shedding light on the pervasive bias issue. The situation was chosen for study because it seemed as close as possible to a case of classic bias; that is, a situation in which someone has the requisite knowledge to answer a question but is unable to access it. Such might be the case when an immigrant who does not know English is asked to take a test in which he/she would know the answer if he/she understood the question.

References

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95–105.

CTBS/McGraw-Hill. (1973). *Comprehensive Test of Basic Skills.* Monterey CA: Author.

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary.* Chicago IL: World Book-Childcraft International.

Elementary School Mathematics Book 2. (1971). Menlo Park CA: Addison-Wesley.

Homan, S. (1980, April). *First application of the Homan Readability Formula, designed to assess individual test items.* Paper presented at the National Council of Measurement in Education Conference, Boston MA.

Hunt, K. W. (1970, February). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development.* Chicago: University of Chicago Press.

Hunter, J. E. (1975, December). *A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items.* Paper presented at the National Institute of Education Conference on Test Bias, MD.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 117–160). Baltimore MD: Johns Hopkins University Press.

Ironson, G. H., & Craig, R. (1982). *Item bias techniques when amount of bias is varied and score differences between groups are present.* (Technical Report No. NIE Grant G-81-0045).

Ironson, G. H., & Subkoviak, M. J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16*(4), 209–225.

Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109–118.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross cultural psychology.* Amsterdam: Swets and Zeitlinger.

Peterson, N. S. (1977, June). *Bias in the selection rule: Bias in the test.* Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980a). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213–233.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980b). A monte carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17,* 1–10.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143–152.

Science Research Associates. (1974). *SRA Elementary Mathematics Program, 4th Grade Level.* Palo Alto CA: Author.

Shepard, L. A. (1981). Identifying bias in test items. In B. F. Green (Ed.), *Issues in testing: Coaching, dis-*

closure and ethnic bias. San Francisco CA: Jossey Bass.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317–375.

Stanford Diagnostic Math Test, Brown Level Form A. (1976). New York: Harcourt Brace Jovanovich.

Subkoviak, M. J., Mack, J. S., Ironson, G. H., & Craig, R. D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21*, 49–58.

Uprichard, A. E. (1979–80). *Pinellas County Math Test Grade 2*. Pinellas County FL.

Uprichard, A. E. (1979–80). *Pinellas County Math Test Grade 4*. Pinellas County FL.

Acknowledgments

The authors thank John Hildebrand and the Hillsborough County School System for their assistance with this project. Special recognition is extended to Jone McFadden, Stephanie Merritt, Rachel Monroe, and Jim Thomas for their help with the data collection.

Author's Address

Send requests for reprints or further information to Susan Homan, EDU 306D, University of South Florida, 4202 Fowler Avenue, Tampa FL 33620, U.S.A.