Implications of the Golden Rule Settlement for Test Construction

Robert L. Linn
and
Fritz Drasgow
University of Illinois at Urbana-Champaign

The authors present the results of an application of Golden Rule procedures to items of the Scholastic Aptitude Test. Using item response theory, their analyses indicate that the Golden Rule procedures are ineffective in detecting biased items and may undermine the reliability and validity of tests.

Before reviewing some implications of widespread applications of the procedures contained in the Golden Rule settlement for the construction and psychometric properties of tests, it is important to emphasize a fact that seems to be overlooked in the courts and by legislators proposing expanded use of the procedures. It is simply that psychological tests *do not measure innate abilities or aptitudes*. Instead, they assess a test taker's current repertoire of knowledge and skills. An individual's repertoire of knowledge and skills is certainly affected by his or her "environment"; consequently, we would expect differences in mean test performance for groups whose environments differ. Family income and parental education are two aspects of "environment" that are known to be related to educational achievement. Because whites and blacks have substantially different mean family incomes and parental educations, we should expect mean differences in test scores. The key point is that unequal environments imply unequal educational achievements and a well-constructed test should reflect this fact.

Background

Classical item analysis techniques have traditionally emphasized two item characteristics: item difficulty (i.e., the proportion of test takers giving the correct answer to an item) and item discriminating power (i.e., the correlation between scores on a given item and total test scores). Although both of these statistics have proven useful in test construction, they have some serious limitations. One of the most notable limitations is that they are highly dependent on the nature of the sample of test takers. Item difficulties based on a well-trained and highly competent sample, for example, would obviously be expected to be quite different from item difficulties based on a less competent sample.

The role of item difficulties in selecting items for a test has recently taken on a new dimension. In No-

---

Robert L. Linn is Professor of Educational Psychology at the University of Illinois at Urbana-Champaign, 1310 S. 6th St., Champaign, IL 61820. He specializes in educational measurement.

Fritz Drasgow is Associate Professor of Psychology at the University of Illinois at Urbana-Champaign, 219D Psychology Building, Champaign, IL 61820. He specializes in psychometrics and industrial psychology.

This article is based on a paper presented at the Annual Meeting of the American Psychological Association, Washington, DC, August 23, 1986 as part of a Division 5 Public Affairs Symposium entitled Golden Rule Revisited. Since the paper was written, Gregory R. Anrig, President of Educational Testing Service, published a statement in the January 1987 issue of *The APA Monitor* in which he stated that he now believes that it was a mistake for ETS to enter into the agreement contained in the Golden Rule settlement. We applaud Mr. Anrig's willingness to consider the unintended consequences of the agreement and his candor in stating that his earlier approval of the settlement was "an error in judgment."

vember 1984, the Illinois Department of Insurance and the Educational Testing Service agreed to an out-of-court settlement with the Golden Rule Insurance Company and five individuals who had failed portions of the Illinois insurance licensing examinations. Although the "Golden Rule" settlement included a number of other provisions, such as the agreement to collect racial and ethnic data on a voluntary basis and the agreement to disclose one form of the test every other year, the provisions for the use of item difficulties for item classification and selection have received the most attention and are the focus of this paper.

Under the provisions of the "Golden Rule" settlement, items are classified into one of two types according to the following definitions:

> Type I—those items for which (a) the correct-answer rates for black Examinees, white Examinees, and all Examinees are not lower than forty percent (40%) at the .05 level of statistical significance, and (b) the correct-answer rates of black Examinees and white Examinees differ by no more than fifteen (15) percentage points at the .05 level of statistical significance; or Type II—all other items. (Golden Rule Insurance Company et al. v. Washburn et al., 1984, p. 10)

The settlement goes on to stipulate that examinations shall be assembled by ETS "in accordance with the subject matter coverage and weighting of the applicable content outline" (Golden Rule, p. 10). However, where possible, tests are to be constructed by using Type I items with preference given to items with smallest between-group differences in correct-answer rates, that is, item difficulties. Type II items may be used only when there are not a sufficient number of Type I items satisfying the constraints of the content outline. When it is necessary to use Type II items, preference is to be given to items for which the correct-answer rates of blacks and whites differ least.

We shall not attempt to evaluate the impact of the Golden Rule settlement on the Illinois insurance licensing examinations, but will simply note that such an evaluation would depend on the test content, the pool of appropriate items, and the characteristics of the populations of persons who take the examinations. The impact on the reliability and validity of these particular examinations may be benign. It should be noted, however, that even if the settlement does not have negative effects on the reliability and validity of these particular examinations, it is based on pragmatics rather than on sound psychometric principles. For this reason it sets a bad precedent, one that could seriously undermine the validity of many other tests to which the procedures might be applied.

Although out-of-court settlements do not normally have the widespread influence that would be expected of a major court decision, the Golden Rule settlement has attracted a great deal of attention and has provided a model for subsequent settlements (e.g., Allen v. the Alabama State Board of Education) and legislation that has been introduced in California and New York. For example, the February 21, 1986 version of the California Assembly Bill introduced by Assembly Member Moore sought to apply the provisions of the Golden Rule settlement to a wide range of tests administered in California including professional licensure examinations and college admissions tests. The bill also expanded the number of groups for which comparisons of item difficulties would be required from two to five: black, Hispanic, Asian, American Indian, and white. Organized support for similar legislation is being provided by the National Center for Fair and Open Testing (Biemiller, 1986).

The potential impact of legislation incorporating the provisions of the Golden Rule settlement can be illustrated by a simple analysis of one test, the Verbal section of the Scholastic Aptitude Test (SAT). A scatterplot of the item difficulties for black and white students on the Verbal section of one form of the SAT can be found in Lord's (1980, Figure 14.2.1, p. 214) book on item response theory. Application of the dual criteria of a minimum of 40% correct for both groups of test takers and a maximum between-group difference of 15 percentage points would have led to the classification of only 25 of the 85 items on that form of the SAT as Type I items. The remaining 60 Type II items could have been used only if there was not a sufficient supply of other pretested items that satisfied the criteria of Type I within the constraints of the content specifications of the test.

Admittedly, the SAT differs from a licensure in a number of respects. It does not have a focus on a clearly defined achievement domain that corresponds to a course of studies designed to prepare individuals to be competent to practice a particular occupation. Nor is the SAT designed to ensure that examinees have the minimum level of knowledge and skills needed to protect the public. However, as was noted above, some proponents of the Golden Rule procedures would like to see them applied not only to tests used for certification and licensure, but to tests such as the SAT that are used in the college admissions process.

The SAT example illustrates that the provisions of the Golden Rule settlement could have a major impact on some tests. By itself, however, it does not provide any indication of the nature of the impact on either the psychometric characteristics of the resulting tests or on the magnitude of the differences in total test scores of the groups used in the classification of items. Nor does it provide any indication of the effect that the application of the procedure would have on possible bias. Each of these three issues will be addressed in the following sections of this paper.

Total Test Score Differences

It seems intuitively reasonable to assume that the elimination of Type II items from a test would result in smaller differences in the group means on the total test score. Such a result would be quite likely if only the second of the dual criteria were used to define Type I items. Application of the first criterion, which requires a minimum correct-answer rate of at least 40% in each group, however, could, in fact, lead to larger rather than smaller between-group differences on the total test score. This is so because, at least on some tests, the difference in item difficulty is often smaller on the more difficult items than on the easier items.

The scatterplot of item difficulties presented by Lord (1980) can again be used to illustrate this point. On

that particular form of the Verbal section of the SAT, most of the items that had the smallest between-group differences in item difficulties would have been relegated to the Type II category because less than 40% of both the white and the black groups of test takers answered the items correctly. For example, a total of 18 of the 85 items had a difference in item difficulty between white and black test takers of .05 or less. That is, the between-group difference is no more than one third as large as the maximum allowed for a Type I item. Such items would presumably be preferred by proponents of the Golden Rule procedure who seek to reduce the magnitude of the difference between white and black test takers on the total test score. However, 17 of the 18 items with the smallest differences in difficulty would actually have been classified as Type II because less than 40% of the black test takers *and* less than 40% of the white test takers answered the items correctly. Elimination of a large number of difficult items with small differences in correct-answer rates on such a test could actually increase rather than decrease the average difference on the total test score.

Bias

Group differences in average performance on tests are a cause for concern and provide one of the motivations for seeking changes along the lines of the Golden Rule procedure. The fact that group differences in average test performance occur consistently on a wide variety of tests also is often taken as prima facie evidence of bias. This popular view of bias contrasts with psychometric definitions of bias and is apparently based on what Anastasi has referred to as a "confusion of measurement and etiology" (1961, p. 389). As stated by Anastasi, "No test score can eliminate causality. Nor can a test score, however derived, reveal the origin of the behavior it reflects. If certain environmental factors influence behavior, they will also influence those samples of behavior covered by tests" (Anastasi, p. 389).

Differences in educational experiences lead to differences in the knowledge and skills that tests are intended to measure, and our society is, unfortunately, a long way from providing an equal educational opportunity to all. On average, some minority groups are less likely to have access to high-quality educational opportunities and consequently are less likely to develop the same level of the knowledge and skills measured by tests. The elimination or artificial reduction of differences in average test scores might conceal this situation, but it would not rectify it.

Although group differences in performance on tests do not necessarily imply bias, the possibility of bias certainly should not be ignored. In rejecting the approach embodied in the Golden Rule settlement, we are not arguing against serious efforts to identify and eliminate sources of bias in tests. The 1985 *Standards for Educational and Psychological Tests* (American Psychological Association, American Educational Research Association, & NCME, 1985) encourages the use of "expert judges both to select item material and to eliminate material likely to be inappropriate or offensive for groups in the test-taking population" (p. 26). The *Standards* also encourages the use of statistical analyses "to detect and eliminate aspects of test design, content, or format that might bias test scores for particular groups" (p. 27). We strongly support these recommendations, but believe that the approach used in the Golden Rule settlement is an inappropriate response to the problem.

If it is accepted that groups can differ in the knowledge or skills that a test is intended to measure, it follows that a difference in the proportion of test takers who answer a particular item correctly is not necessarily an indication that the item is biased. An adequate approach to detecting items that introduce irrelevant sources of difficulty for members of a particular group requires a means of distinguishing between differences that are due to group differences in the developed skills of the test takers and those that are due to extraneous factors. The most widely accepted psychometric approach to this problem is based on item response theory.

According to item response theory, the probability of a correct answer to an item depends only on the underlying skill of the test taker and one or more item parameters.

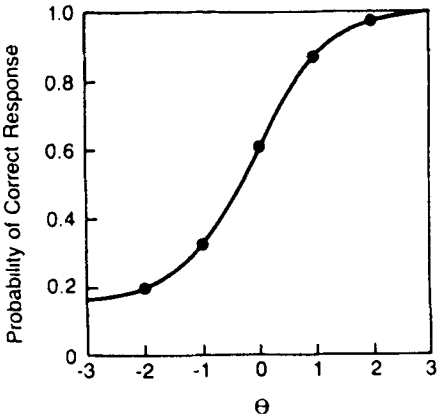

FIGURE 1. *Illustration of an item response curve*

An illustration of an item response curve is provided in Figure 1. The response curve is simply a plot of the probability of a correct response as a function of the attribute measured by the test. The Greek letter theta (θ) is usually used to denote the attribute. Note that individuals with high θs are expected to answer correctly with high probability and individuals with low θs are expected to answer correctly with low probability.

Item response curves lead naturally to a definition of item bias. An item is considered to be biased if the item response curves "estimated in samples from different subpopulations are not identical within the limits of sampling fluctuations" (Hulin, Drasgow, & Parsons, 1983, p. 167).

Several authors (e.g., Camilli & Shepard, in press; Drasgow, 1987; Hunter, 1975; Lord, 1977, 1980) have used item response theory to demonstrate the inadequacies of item difficulties as indices of item bias. By assuming that an item is unbiased according to the above item response theory definition, it can easily be shown that a variety of differences in item difficulty can be obtained for two groups that differ in the underlying attribute measured by the test. Figures 2 and 3 provide two such illustrations. In both figures, the ability distribution of each group is assumed to be normal with a standard deviation of 1.0. The Group 1 mean is $-.5$ and the Group 2 mean is .5. That is, the mean ability differs by 1 within-group standard deviation, a figure roughly comparable to that often encountered in practice for white and black test takers.

The item response curves in both

figures are assumed to be identical for both groups and follow the 3-parameter logistic model, that is, the items are unbiased according to the item response theory definition. Both items have lower asymptotes equal to .2 (i.e., c = .2), and location parameters halfway between the group ability means (i.e., b = 0). The items differ only in the values of their discrimination parameters, with a = 1 for the item in Figure 2 and a = .5 for the item in Figure 3. As can be seen in the figures, the difference in proportion correct is .22 for the item in Figure 2 and .14 for the one in Figure 3. Thus, Item 2, the less discriminating item, would satisfy the criteria for a Type I item and therefore be preferred to Item 1, which fails to meet the criteria of a Type I item.

The simple comparison of the results in Figures 2 and 3 illustrates a general finding. The difference in item difficulty depends on the magnitude of the a parameter. In particular, for given differences in θ distributions for two groups, given values of the b and c parameters, and with b between the mean θs for the two groups, the difference in item difficulty increases as a function of a. In other words, the better the item in terms of its discriminating power, the more likely it is to show a large difference in item difficulty and therefore be classified as a Type II item according to the Golden Rule procedure.

The difference also depends on the other two item parameters and on the actual difference in θ between the two groups. Table 1 lists differences in item difficulties for selected values of the a, b, and c parameters based on the same assumed distributions of θ that were used in Figures 2 and 3. As before, the results in Table 1 are based on the 3-parameter logistic model, and all the items are assumed to be unbiased according to the item response theory definition. An inspection of the results in Table 1 reveals three general trends:

1. For a given a and b, the difference in item difficulty decreases as c increases.
2. For a given a and c, the difference in item difficulty decreases as the value of b departs from the midpoint of the group means.
3. For a given b and c, the difference in item difficulty decreases as a decreases.

All three of these trends run exactly counter to the properties required for an item to provide valid measurement. These results support the following conclusion. When items are in fact unbiased, the application of the Golden Rule procedure is more likely to eliminate psychometrically desirable items than psychometrically undesirable items. In fact, if all test takers respond randomly to an item (i.e, the item pro-

TABLE 1

Differences in Item Difficulties for Two Groups With the Same Item Response Curves as a Function of the Item Parameters

	c = 0		
	b = −1.0	b = 0	b = 1.0
a	$p_2 - p_1$	$p_2 - p_1$	$p_2 - p_1$
.25	.09	.10	.09
.50	.14	.18	.14
.75	.15	.24	.15
1.00	.16	.28	.16
1.25	.16	.30	.16
1.50	.15	.32	.15
1.75	.15	.33	.15
2.00	.15	.34	.15
	c = .2		
.25	.07	.08	.07
.50	.11	.15	.11
.75	.12	.19	.12
1.00	.13	.22	.13
1.25	.13	.24	.13
1.50	.12	.26	.12
1.75	.12	.27	.12
2.00	.12	.27	.12

Note. Normally distributed ability, means, and standard deviations equal −.5 and 1.0 for Group 1 and .5 and 1.0 for Group 2.

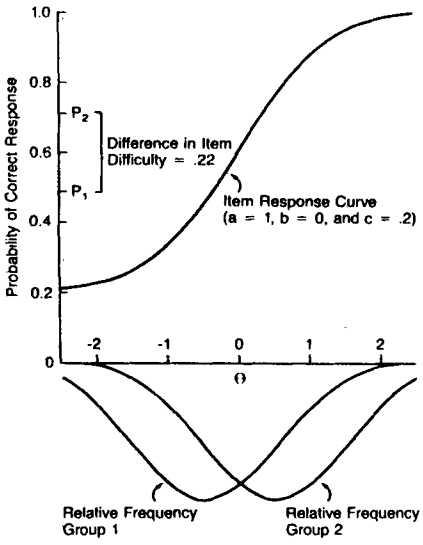

FIGURE 2. *Illustration of item difficulties for two groups with the same item response curve (Item 1)*

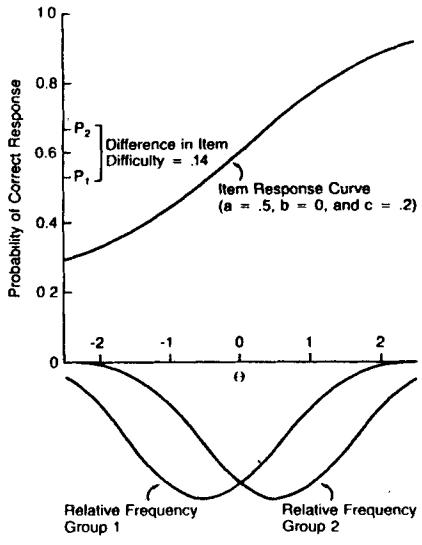

FIGURE 3. *Illustration of item difficulties for two groups with the same item response curve (Item 2)*

vides no valid information about the attribute being measured by the test), it will show no difference in item difficulties and thus satisfy the Golden Rule 15% criterion. Of course, such an item should not be included on a test.

The above analysis is based on items that are unbiased in the sense of item response theory. Here an item is defined as unbiased if test takers from different groups (say, white and black) have equal probabilities of a correct response when they have equal standings on the attribute measured by the test (i.e., equal θs). Of course, it is important to determine whether items on a test are unbiased. Hence, it is relevant to ask how well the Golden Rule procedure would work in identifying items that are biased according to the item response theory definition (i.e., test takers with equal θs have unequal chances of correct responses). A recent paper by Camilli and Shepard (in press) provides convincing evidence that not only is the use of differences in item difficulties as indices of item bias flawed because it misclassifies unbiased items as biased (Type I errors), as was shown above and in several earlier papers (e. g., Hunter, 1975; Lord, 1977, 1980), but the approach is also flawed because it is insensitive to bias when it does exist (Type II errors).

A number of theoretically sounder and more powerful techniques for detecting items that are biased are available. The statistical techniques based on item response theory that have been used, for example, by Shepard, Camilli, and Williams (1984, 1985) are much to be preferred to the simplistic approach implicit in the Golden Rule settlement. The Mantel-Haenszel procedure that has recently been introduced by Holland and his colleagues (e.g., Holland, in press; Holland & Thayer, 1986) provides another powerful alternative. We encourage the continued investigation and use of these techniques.

Validity

The Golden Rule procedure threatens to undermine the most important characteristics of sound tests. Application of the procedure is likely to reduce reliability. More importantly, it is likely to degrade the validity of tests. It will reduce reliability because it will favor items with poor discriminating power. It will degrade validity because it is precisely those items that provide the best measurement of the underlying attribute being measured that are most likely to fail to meet the criteria of a Type I item.

> Furthermore, the use of these statistical criteria is likely to distort the construct validity [of the test] because the criteria are apt to favor items requiring lower-level associational processes over items requiring higher-level abstractions and concepts. Though written in a different context, Anderson's (1972, p. 165) conclusion that an overreliance on item statistics "tortures validity" provides an apt description in the present context. (Linn, 1986, p. 81).

Some of the concerns that have led to the support for the Golden Rule procedure are legitimate. The possibility of item bias deserves continued attention. However, this needs to be done using more justifiable techniques.

The use and misuses of test results for minority students also need to be constantly reviewed. In this regard, it is crucial to keep in mind that a test can only measure performance at a given point in time. It cannot reveal an innate capacity. If groups differ in the quantity and quality of their educational experiences, both in and out of school, it is reasonable to expect that those differences will influence test scores. Keeping such differences and the effects of previous discrimination in mind is in keeping with the spirit of recent Supreme Court decisions on affirmative action. Application of procedures such as those included in the Golden Rule agreement, however, will do more to hide these problems than to solve them.

References

Allen v. Alabama State Board of Education, 81-697-n (consent decree filed with United States District Court for the Middle District of Alabama Northern Division, 1985).

American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1961). Psychological tests: Uses and abuses. *Teachers College Record, 62,* 389–393.

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42,* 145–170.

Anrig, G. R. (1987, January). "Golden Rule": Second thoughts. *APA Monitor,* p. 3.

Biemiller, L. (1986, January 8). Critics plan assault on admissions tests and other standard exams. *The Chronicle of Higher Education,* p. 4.

Camilli, G., & Shepard, L. A. (in press). The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics.*

Drasgow, F. (1987). A study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Golden Rule Insurance Company et al. v. Washburn et al., 419-76 (stipulation for dismissal and order dismissing case, filed in the Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL, 1984).

Holland, P. W., (in press). On the study of differential item performance without IRT. *Proceedings of the Military Testing Association.*

Holland, P. W., & Thayer, D. T. (1986, April). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

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 items. Paper presented at the National Institute of Education Conference on Test Bias. Annapolis, MD.

Linn, R. L. (1986). Bias in college admissions. In *Measures in the college admissions process: A College Board colloquium* (pp. 80–86). New York: The College Board.

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* (pp. 19–29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93–128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77–105.