Journal of Educational Statistics
Spring 1987, Vol. 12, No. 1, pp. 87–99

The Inadequacy of ANOVA for Detecting Test Bias

Gregory Camilli
Rutgers University
and
Lorrie A. Shepard
University of Colorado, Boulder

Key words: *test bias, item bias, ANOVA, group-by-item interaction*

The inadequacy of ANOVA for detecting bias in test items should already be well understood, yet it persists as a popular method. Here, previous arguments are extended to explain why ANOVA may obscure test bias when it exists, as well as create false impressions of bias. In fact, it is demonstrated in this paper that ANOVA will fail to detect even absurdly large amounts of bias. More specifically, it is shown that bias contributes relatively more to the group main effect than to the group-by-item interaction.

Analysis of variance (ANOVA) is one of the most widely used procedures for assessing group bias in the internal structure of a test. According to this method, bias is defined as a significant group-by-item interaction indicating that some test items are relatively more difficult for one group than another. ANOVA was used in early studies of item bias (Cleary & Hilton, 1968) and continues to be recommended as a bias detection technique (Plake, 1981; Plake & Hoover, 1979).

Jensen (1980) reviewed existing studies and concluded that there was no evidence of internal bias in standardized tests of mental ability. The lion's share of these internal studies relied on ANOVA or highly similar methods based on item difficulty differences. Looking over the same body of work, Gordon and Rudert (1979) likewise concluded that IQ tests are not culturally biased. They further asserted that the race-by-item interaction method is a "powerful one," quite capable of detecting questionable items when they are present (p. 179). Gordon and Rudert went on to say that absence of race-by-item interactions in all of these studies places a severe constraint on the differential experience hypothesis espoused by most test critics. Both Jensen (1984) and Gordon (in press) contend that the only kind of bias that could exist but remain undetected by the interaction term would be some implausible kind of constant bias that would affect all items in precisely the same way.

The inadequacy of ANOVA for detecting internal test bias should be well known. Hunter (1975) and Lord (1977) demonstrated the fallacy in using differential item difficulties (p-values) as indices of bias. ANOVA provides nothing more than an omnibus test of item p-value differences. However, when there are real differences in performance between two groups, the magnitude of item p-value differences will vary as a function of item discrimination. Thus, highly discriminating items will appear to be biased simply because they better distinguish between groups, that is, show a larger difference in percentage correct. The varying p-value differences give rise to an apparent race-by-item interaction. The deficiencies of classical p-value methods will be recapitulated in the first section of the paper.

Unfortunately, the arguments against ANOVA and other p-value methods have enjoyed a popular one-sided interpretation: namely, the appreciation that p-value methods could create false instances of bias. There is no evidence of an equal concern that these methods might obscure or miss real incidents of bias. For example, Jensen (1984) added the following discussion of artifactual effects to his interpretation of group-by-item interactions:

> The observed group × item interaction, in virtually all cases that we have examined, turns out to be an artifact of the method of scaling item difficulty. Essentially, it is a result of the nonlinearity of the item-characteristic curve. As I failed to explain this artifact adequately in my treatment of the group × item method in *Bias in Mental Testing*, I will attempt to do so here. (p. 536). . . .
>
> The practical implication of this demonstration for all data that now exist regarding group × item interaction is that the small but significant observed group × item interactions would virtually be reduced to nonsignificance if the artifact due to ICC nonlinearity were taken into account. It is likely that the correct conclusion is that in most widely used standard tests administered to any American-born English-speaking populations, regardless of race or ethnic background, group × item interaction is either trivially small or a nonexistent phenomenon. (pp. 537–538)

The purpose of the present paper is to explain more clearly the inadequacy of the ANOVA method for detecting internal test bias. Especially, it is important to understand how p-value methods will fail to detect real occurrences of bias. In fact, given plausible group differences and amounts of bias, the differential difficulty (bias) contributes more to the between-groups effect than to the interaction. We first offer a heuristic demonstration, then an algebraic demonstration, and finally a simulation study.

Conceptual Argument

Hunter (1975) used item characteristic curves to examine bias methods proposed by Green and Draper (1972), Angoff (1972), and Jensen (1974). Item characteristic curves (ICCs) are mathematical functions that relate the

probability of a correct response on an item to the underlying continuum of achievement. Hunter adopted examples where test items were all unbiased, that is, the ICCs for two groups were the same. Equivalence of ICCs defines unbiasedness since it implies that the probability of answering an item correctly is the same regardless of group membership.

Three items varying in difficulty are shown in Figure 1[1] (adapted from Hunter, 1975). The mean differences on each of these items are designated

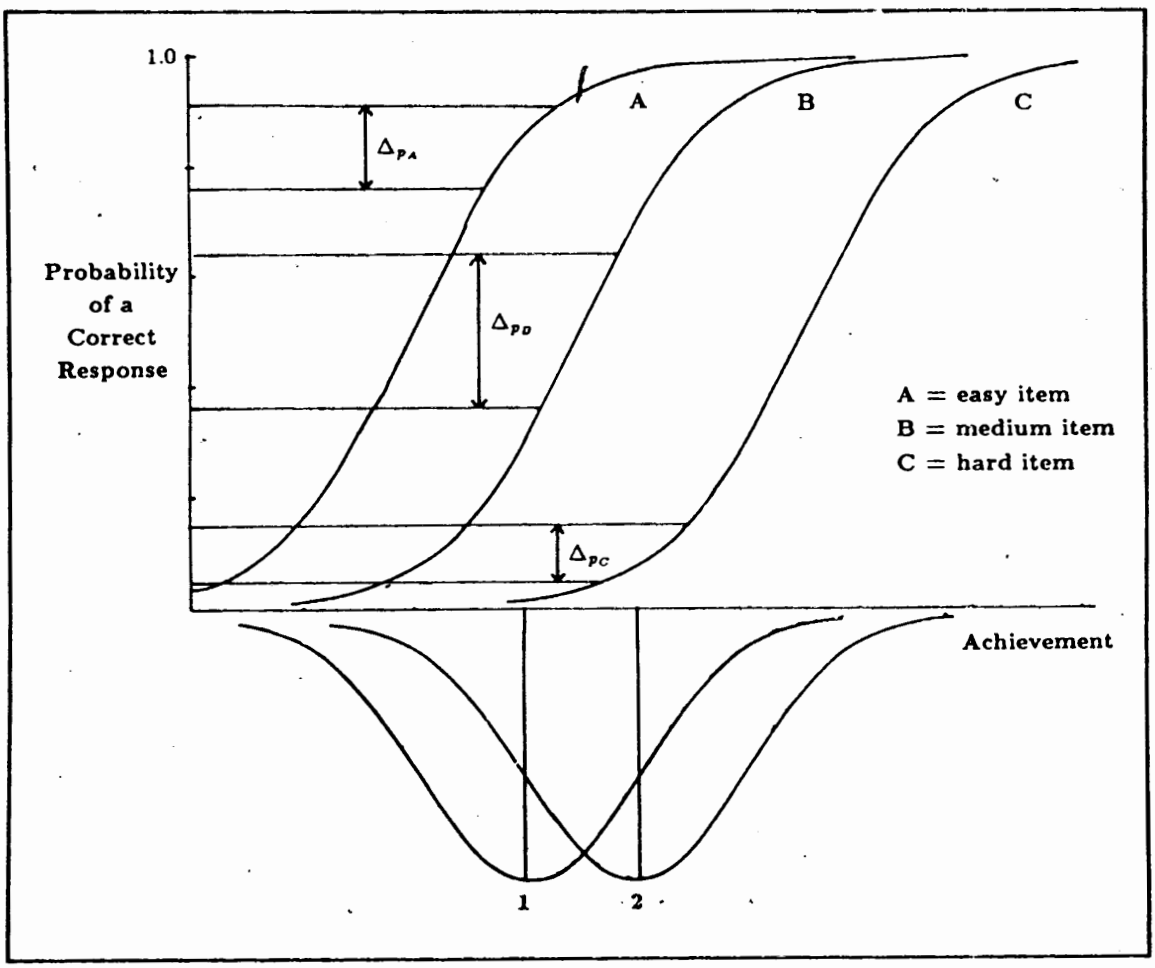

FIGURE 1. *Unbiased situation; the item means for an easy, medium, and hard item for each of two groups with different average achievement levels (from Hunter, 1975)*

[1] In Figures 1 and 2, note that the p-values for each group do not fall precisely at the intersection of the group mean and the ICC. Except when the ICC is centered over the group mean (i.e., $\bar{\theta} = b$), the actual p-value will be slightly closer to .5 than is $P(\bar{\theta})$. Classical p-values were obtained by the approximation $\sum_{i=1}^{13} f_i P(\theta_i)$, where f_i is the proportion of area in the ith interval according to the logistic density function.

for two groups that differ in average achievement. Item A is an easy item and has a high probability of being answered correctly in both groups. Thus, the group difference is small for Item A. Similarly, Item C is very difficult and yields a small difference. Item B is centered between the two group means and produces a much larger p-value difference. Because of the greater p-value difference, Item B would appear to be biased by the Angoff delta plot method. Likewise, the non-uniform p-value differences would result in a significant group-by-item interaction. Thus, Hunter demonstrated that artifactual bias may arise merely as a function of the mean group difference in a situation where the ICCs were in fact identical for both groups.

Now let us consider a situation where bias is present. We will examine easy, medium, and difficult items in separate graphs. In Figure 2a, 2b, and 2c, the dotted curves show the biased response functions for the lower scoring group. These curves are shifted to the right, indicating that the item difficulty is relatively greater, or the probability of answering correctly lower, for a given location on the achievement dimension. Corresponding p-value differences are shown between the Group 2 means on the solid curve and the Group 1 means on the dotted curve.

In the Figure 2 example, Items A and B have large p-value differences. The amount of bias chosen for illustration corresponds to the ICC differences observed in a real-data study (Shepard, Camilli, & Williams, 1984). The resulting p-value difference for Items A and B is about the same magnitude as the artifactual "bias" in Figure 1. In this example, Item C has a much smaller p-value difference, although the actual bias is the same. *The amount of bias measured by differential difficulty is more a function of item location than of real shifts in the ICCs.* If a test were comprised mostly of items like A and B in Figure 2, items like C would appear to be biased *against the high scoring group* (although the reverse is true) simply because the p-value difference is less than average.

If we had a three-item test comprised of the three items in Figure 2, a significant group-by-item interaction would be obtained because Item C is markedly different from Items A and B. The example contrived in Figure 2 does not depend, however, on having created constant bias in the three items. If only Item A or only Item B were biased, an interaction would be detected of the same magnitude as when all three items were biased. These interactions would also be of approximately the same size as the artifactual interaction in Figure 1. If only Item C were biased, the interaction would be virtually the same as when no item was biased.

Although these examples do not constitute exhaustive proof that ANOVA is inadequate for bias detection, it should be disquieting that real bias will sometimes be ignored by the interaction term. Furthermore, when bias does produce larger p-value differences, they may be of the same magnitude as artifactual differences created by item discrimination.

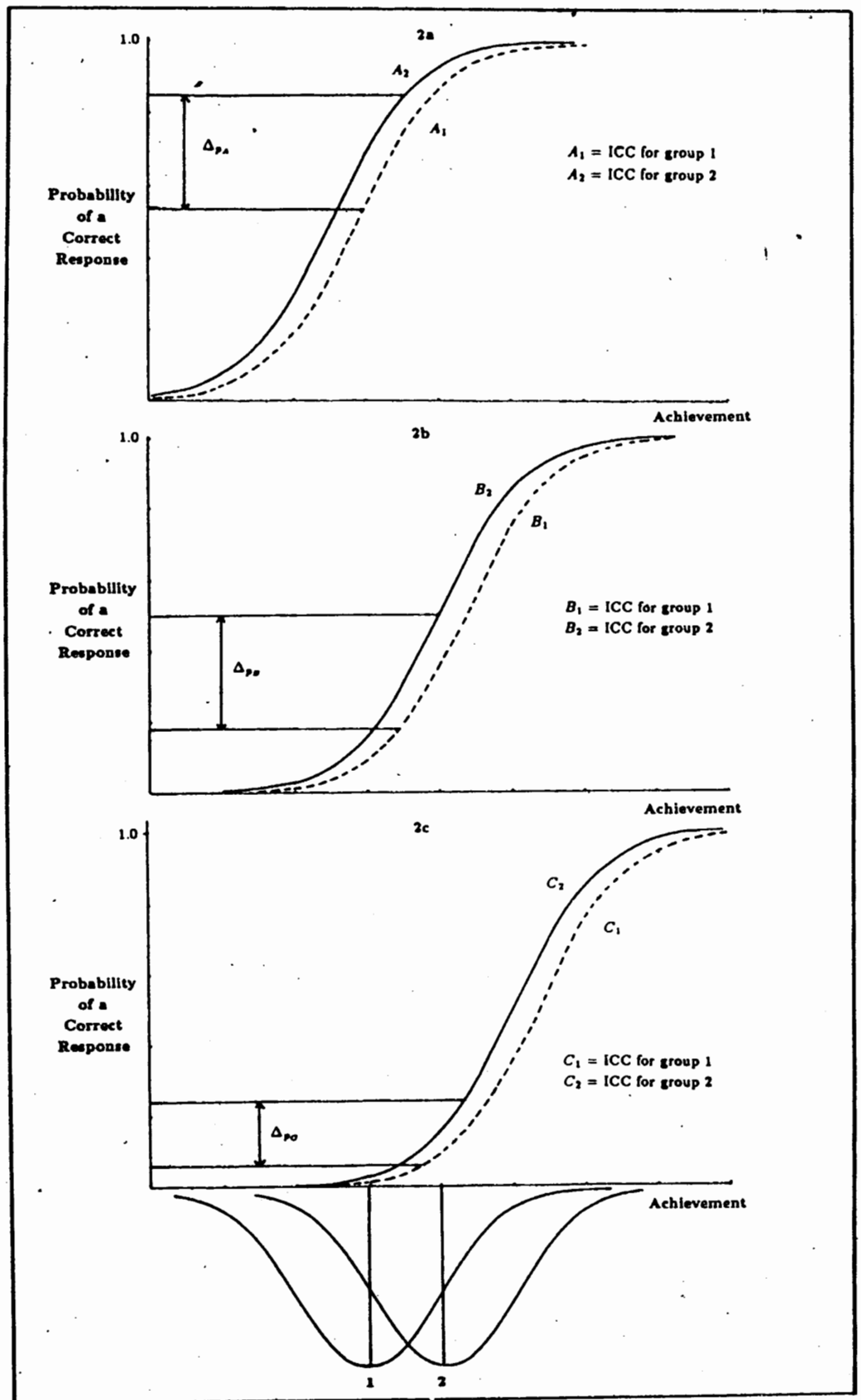

FIGURE 2. *Biased situation; item means for (a) easy, (b) medium, and (c) hard biased items for two groups with different average achievement levels*

Whether a "significant" interaction will occur depends on the item difficulties in relation to the group means.

Algebraic Demonstration

Item response curves are inherently nonlinear; thus questions concerning the applicability of ANOVA techniques naturally arise. But it is not only this nonlinearity that is the basic weakness of the ANOVA method for detecting test bias. To see this, consider a test that is adequately represented by a one-parameter (Rasch) test model:

$$P_j(\theta) = 1/[1 + exp\{-1.7(\theta - b_j)\}]$$

where P_j is the probability of correctly responding to item j, θ is the ability level of an examinee, and b_j is the difficulty of item j.

The problems with nonlinearity in theory can be eliminated mathematically by making the substitution $X = log[P(\theta)/Q(\theta)]$ for the observed item response, that is, wrong (zero) or right (one). The expected logit score, μ, is a simple linear function of θ and item difficulty, namely,

$$\mu_{ij} = \bar{\theta}_i - b_j,$$

where i denotes group and j denotes item.

We note that X is an idealized or true item response score that is never observed. This allows the following examination of the ANOVA bias technique based on a highly simplified example and performed with logit scores. Thus we arrive at the question, "How good is ANOVA at detecting bias in the ideal case where there is no confounding effect due to nonlinearity?"

For this hypothetical analysis, suppose there are three items, one of which is biased. Suppose also that there is a true mean difference between two groups (say Groups 1 and 2) of α logits. The expected difference in mean performance of the two groups on the two unbiased items is then also α logits because

$$\mu_{2j} - \mu_{1j} = (\bar{\theta}_2 - b_j) - (\bar{\theta}_1 - b_j) = \bar{\theta}_2 - \bar{\theta}_1 = \alpha.$$

On the third item that is biased suppose that

$$\mu_{23} - \mu_{13} = \bar{\theta}_2 - \bar{\theta}_1 + \beta = \alpha + \beta.$$

These data would permit a subject-within-group-by-item repeated measures ANOVA. We are interested in examining two sources of variation from this design and the relative effect of the magnitude of bias, β; for simplicity we assume "items" is a fixed factor. The first source is the variance component for groups that is given by

$$\sigma_G^2 = \sum_{i=1}^{2} (\mu_{i\cdot} - \mu_{\cdot\cdot})^2 = 1/2(\mu_{1\cdot} - \mu_{2\cdot})^2.$$

The second source of variation is the variance component for the group-by-item interaction that is given by

$$\sigma^2_{G\cdot item} = 1/2 \sum_{i=1}^{2} \sum_{j=1}^{3} (\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot})^2$$

where j is the subscript for items.

To derive numerical estimates from these formulae we note that the group differences for each item can be expressed as

item	difference
1	$\mu_{21} - \mu_{11} = \alpha$
2	$\mu_{22} - \mu_{12} = \alpha$
3	$\mu_{23} - \mu_{13} = \alpha + \beta$
Total mean	$\mu_{2\cdot} - \mu_{1\cdot} = \alpha + \beta/3.$

So the σ^2_G can be rewritten

$$\begin{aligned}\sigma^2_G &= 1/2(\mu_{2\cdot} - \mu_{1\cdot})^2 \\ &= 1/2(\alpha + \beta/3)^2 \\ &= 1/2(\alpha^2 + 2/3\alpha\beta + 1/9\beta^2).\end{aligned}$$

The variance component for the 2×3 group-by-item interaction can be found by deriving each of the six $\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}$ terms. For example, for $i = 1$ and $j = 1$ we have

$$\begin{aligned}\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} &= \mu_{11} - (\mu_{11} + \mu_{12} + \mu_{13})/3 - (\mu_{11} + \mu_{21})/2 \\ &\quad + (\mu_{11} + \mu_{12} + \mu_{13} + \mu_{21} + \mu_{22} + \mu_{23})/6 \\ &= 1/6[-2(\mu_{21} - \mu_{11}) + (\mu_{22} - \mu_{12}) + (\mu_{23} - \mu_{13})] \\ &= 1/6[-2\alpha + \alpha + (\alpha + \beta)] \\ &= 1/6\beta.\end{aligned}$$

Repeating these calculations for each cell gives

Cell	$\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}$
11	$+1/6\beta$
12	$+1/6\beta$
13	$-2/6\beta$
21	$-1/6\beta$
22	$-1/6\beta$
23	$+2/6\beta.$

Squaring and summing across cells gives

$$\sigma^2_{G\cdot item} = 1/2(12/36)\beta^2 = 1/6\beta^2.$$

To understand the relative capability of the interaction term to reflect the bias effect, β, we compare the interaction variance component to the between-groups variance component. The *increase* in the group variance component ($+\sigma^2_G$) because of β equals

$$(1/2)(\alpha + \beta/3)^2 - (1/2)\alpha^2 = (1/6)(2\alpha + \beta/3)\beta.$$

The *increase* in the group-by-item variance component ($+\sigma^2_{G\cdot item}$) because of β is given by

$$(1/6)\beta^2.$$

The conditions under which β will contribute more to σ^2_G than to $\sigma^2_{G\cdot item}$ can be determined by solving the following inequality

$$+\sigma^2_G / +\sigma^2_{G\cdot item} > 1$$

or

$$\frac{(1/6)(2\alpha + \beta/3)\beta}{(1/6)\beta^2} = \frac{2\alpha + \beta/3}{\beta} > 1.$$

Thus, when $\alpha > 1/3\beta$, the bias effect will contribute more to the between-groups effect than to the group-by-item interaction. From previous studies we know that the group mean difference, α, far exceeds the typical amount of bias, β (Shepard, Camilli, & Williams, 1984). Therefore, plausible amounts of bias will contribute more to the group effect rather than be detected by ANOVA as bias.

Simulation Study

Data Generation

Two population groups were specified with normally distributed θ abilities. The higher scoring group, corresponding to the majority group in typical black-white comparisons, was set to have a mean of 0.0 and a standard deviation of 1.0. The lower scoring group was set to have a mean of -1.0 and standard deviation of .75. When these data were combined for two groups of equal size, the pooled standard deviation would be expected to be .884; the between-group effect (i.e., the difference between means divided by the pooled standard deviation) would be 1.13 in these units.

Four different unbiased tests were created by varying the location of item difficulties. In each case, 40 items were roughly normally distributed within a specified range. For a very easy test, item *b*'s ranged from -3 to 0 on the θ scale. One test, "centered" in the region of the two group means, covered the range from -2 to 1; another "narrow centered" test covered the range from only -1 to 0. Finally, a very hard test was created with item *b*'s ranging from -1 to 2.

Different levels of bias against the low scoring group were then simulated by decreasing the difficulty of items for the high-scoring group for 1/4 or 1/2 of the items. In addition, the amount of difficulty shift (Δb) was set at five different levels: .35, .50, .75, 1.00, and 1.50. Two proportions of biased items crossed with five different amounts of bias produced 10 bias conditions in addition to the original unbiased test.

Following the above design specifications, item responses were generated using a one-parameter Rasch model:

$$P(\theta) = 1/[1 + exp\{-1.7(\theta - b)\}].$$

To simulate the 0 or 1 item responses of examinees from the two groups, $P(\theta)$ was computed for each examinee on each item. Then a uniform random deviate, U, was drawn from the interval $[0, 1]$. If $U \leq P(\theta)$, a correct response, 1, was recorded. If $U > P(\theta)$, an incorrect response, 0, was recorded. Thus, a binary item score that included some error was used rather than a logit. For each test and bias condition, data were generated for 1,000 examinees, 500 in each group.

ANOVA Results

Forty-four separate data sets were subjected to analysis of variance using a subject-within-group-by-item repeated measures design. Both items and persons were treated as random effects while groups was fixed in contrast to the algebraic example. Variance components (as percentage of total variance) and the obtained group effect are reported in Table 1 for each condition.

The first row for each test is the unbiased condition. The ability of the tests to recover the true difference in groups, $\delta = 1.13$, is shown by the estimated group effect ($\hat{\delta}_G$). As one would expect, the centered-wide test is the most accurate in this respect with $\hat{\delta}_G = 1.16$. The group-by-item interaction accounts for 1% of the variance in the very easy and the very hard unbiased tests. Artifactual bias occurs in the extreme tests since in each case substantial numbers of items are split between the non-discriminating θ range and the highly discriminating θ range (which spans the two means). As expected, the variance accounted for by the between-groups effect is least on the easy test and greatest on the highly discriminating narrow-centered test.

The bias conditions allow us to observe the size of the interaction variance components when bias is modest or severe. It is also possible to compare the relative contribution of the bias effect to between-groups variance versus the interaction. Both the between-groups and the interaction variances increase directly in response to the built-in bias. However, the increase in the group variance is substantially more than the increase in the interaction components. For all four tests the variance due to items remains nearly constant across the bias conditions. As bias goes from zero to very extreme, there is naturally a slight relative decline in both the person-within-group and person-by-item-within–group variances.

On the centered-wide test the interaction accounts for only 5% of the variance even when an absurdly large amount of bias has been built into the test, that is, half of the items have been made easier for the high-scoring group by an extra 1.5 θ units beyond the usual group difference. At the same time, the variance due to groups has increased from 13% to 27% of

TABLE 1

ANOVA variance components and between-groups effect for four simulated tests with 11 conditions of bias

Item difficulty range	Simulated bias conditions: Bias in b	Simulated bias conditions: % biased items	$\hat{\sigma}^2_G$	$\hat{\sigma}^2_{item}$	$\hat{\sigma}^2_{p(G)}$	$\hat{\sigma}^2_{G\cdot item}$	$\hat{\sigma}^2_{i\cdot p(G)}$	$\hat{\delta}_G$
Very easy	0.00	00%	.08	.16	.15	.01	.60	−1.06
	0.35	25%	.09	.16	.14	.02	.60	−1.15
	0.35	50%	.11	.15	.14	.02	.58	−1.25
	0.50	25%	.10	.16	.14	.02	.58	−1.19
	0.50	50%	.12	.15	.13	.03	.57	−1.33
	0.75	25%	.10	.16	.14	.02	.58	−1.23
	0.75	50%	.13	.15	.13	.03	.56	−1.43
	1.00	25%	.11	.16	.14	.03	.57	−1.27
	1.00	50%	.14	.15	.12	.04	.55	−1.52
	1.50	25%	.11	.16	.13	.03	.57	−1.33
	1.50	50%	.16	.15	.11	.05	.53	−1.66
Centered-Wide	0.00	00%	.13	.11	.19	.00	.58	−1.16
	0.35	25%	.15	.11	.18	.01	.56	−1.27
	0.35	50%	.17	.10	.17	.01	.55	−1.38
	0.50	25%	.15	.11	.18	.01	.55	−1.33
	0.50	50%	.18	.10	.17	.01	.54	−1.49
	0.75	25%	.17	.11	.17	.02	.54	−1.41
	0.75	50%	.21	.10	.16	.02	.52	−1.64
	1.00	25%	.18	.11	.16	.02	.53	−1.48
	1.00	50%	.23	.10	.15	.03	.50	−1.79
	1.50	25%	.19	.11	.15	.04	.51	−1.63
	1.50	50%	.27	.09	.12	.05	.46	−2.11
Centered-Narrow	0.00	00%	.16	.02	.22	.00	.61	−1.21
	0.35	25%	.18	.02	.21	.00	.59	−1.32
	0.35	50%	.21	.02	.20	.00	.57	−1.44
	0.50	25%	.19	.02	.21	.00	.58	−1.36
	0.50	50%	.23	.02	.19	.01	.56	−1.53
	0.75	25%	.20	.02	.20	.01	.57	−1.44
	0.75	50%	.25	.02	.18	.02	.54	−1.70
	1.00	25%	.22	.02	.19	.02	.56	−1.52
	1.00	50%	.28	.02	.16	.02	.51	−1.87
	1.50	25%	.23	.03	.17	.03	.54	−1.66
	1.50	50%	.32	.03	.14	.04	.48	−2.16

TABLE 1 (*continued*)

Item difficulty range	Simulated bias conditions: Bias in b	Simulated bias conditions: % biased items	$\hat{\sigma}^2_G$	$\hat{\sigma}^2_{item}$	$\hat{\sigma}^2_{p(G)}$	$\hat{\sigma}^2_{G \cdot item}$	$\hat{\sigma}^2_{i \cdot p(G)}$	$\hat{\delta}_G$
Very hard	0.00	00%	.12	.10	.17	.01	.61	−1.18
	0.35	25%	.13	.09	.17	.02	.59	−1.27
	0.35	50%	.15	.09	.16	.02	.57	−1.38
	0.50	25%	.14	.10	.16	.02	.58	−1.32
	0.50	50%	.17	.09	.16	.02	.56	−1.47
	0.75	25%	.16	.10	.16	.03	.56	−1.41
	0.75	50%	.20	.09	.15	.03	.52	−1.65
	1.00	25%	.17	.07	.15	.04	.54	−1.50
	1.00	50%	.23	.09	.14	.05	.49	−1.84
	1.50	25%	.19	.10	.14	.06	.51	−1.69
	1.50	50%	.29	.09	.12	.07	.43	−2.25

the total variance. The bias has caused the group mean difference to nearly double, increasing from 1.16 to 2.11 standard deviations.

Yet this false group difference would be interpreted as largely a real difference since the group variance is more than five times as large as the interaction variance. For much smaller and more plausible amounts of bias, for example, a .35 shift in 1/4 or 1/2 of the items, the group-by-item interaction contributes only 1% to the total variance. Meanwhile, the increment in the between-groups variance is two or four times as great, increasing from 13% to 15% and 17% for tests with 1/4 and 1/2 of the items biased, respectively. The small amount of bias in 1/4 of the items was enough to increase the difference between groups from 1.16 to 1.27 standard deviations. This change is a 9% increase but would be dismissed as trivial because the interaction variance is only 1%. Interestingly, a change in the amount of group difference on the same order was found on a real math test by Shepard, Camilli, and Williams (1984). In that study items were identified as biased using cross-validated item response theory methods. When seven items biased against blacks were removed from the 32-item test, the mean difference between blacks and whites was reduced from .91σ to .81σ. The bias amount in these items had been small, that is, $b_W - b_B$ from −.20 to −.35. For all of the simulated examples built-in bias adds more to the between-groups variance than to the interaction variance. Furthermore, in an absolute sense the magnitude of the bias has to be of egregious size before it accounts for 5% or more of the variance.

Conclusion

When there is a true difference in group achievement levels, the ANOVA interaction term is incapable of detecting bias that adds or subtracts from this true difference. In this study, the simplistic algebraic demonstration is borne out by the simulation examples. In the presence of group differences, bias against the low-scoring group adds to the between-groups variance rather than creating a group-by-item interaction of practical importance. Both the graphic examples and the different simulation tests illustrate that the location of the items relative to the group means influences both the size of the interaction and the observed mean difference. We do not offer these arguments as proofs that analysis of variance will be inaccurate for bias detection when group means are the same. Our three-item algebraic example should be extended to include multiple biased items, but apparently the mean difference does not have to be very large relative to bias to preclude detection of the bias. Even if two groups are similar, as is the case in some sex comparisons, it is wrong to have to assume what is essentially the object of inquiry, that is, whether the observed difference is real or the result of bias.

The limitations of ANOVA for bias detection should be obvious from the general arguments against classical test theory offered by Lord (1980) and Wright (1977). But apparently these arguments have not been understood, or existing studies that find interactions of only 1% or 2% would not be used to argue with certainty that tests are unbiased. In our most extreme example, the centered-narrow test, bias could account for up to 35% of the mean difference and still explain only 2% of the variance. The specific arguments of Hunter (1975) against several classical bias procedures have been acknowledged but only to the extent that they create artifactual bias. That these methods will also obscure bias has not been appreciated.

Of course, our conceptual and simulated analyses do not prove that bias exists in current standardized tests of ability, only that ANOVA cannot be used to address this issue. Existing bias studies based on ANOVA should be disregarded. And, ANOVA should no longer be recommended as a bias procedure, even for preliminary screening of items.

References

Angoff, W. H. (1972, September). *A technique for the investigation of cultural differences.* Paper presented at the annual meeting of the American Psychological Association, Honolulu. (ERIC Document Reproduction Service No. ED 069 686)

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28,* 61–75.

Gordon, R. A. (in press). Jensen's contributions concerning test bias: A contextual view. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy.* London: Falmer Press.

Gordon, R. A., & Rudert, E. E. (1979). Bad news concerning IQ tests. *Sociology of Education, 52,* 174–190.

Green, D. R., & Draper, J. F. (1972, September). *Exploratory studies of bias in achievement tests.* Paper presented at the annual meeting of the American Psychological Association, Honolulu.

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.

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 40,* 185–244.

Jensen, A. R. (1980). *Bias in mental testing.* New York: The Free Press.

Jensen, A. R. (1984). Test bias: Concepts and criticisms. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 507–586). New York: Plenum Press.

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.

Plake, B. S. (1981). An ANOVA methodology to identify biased test items that takes instructional level into account. *Educational and Psychological Measurement, 41,* 365–368.

Plake, B. S., & Hoover, H. D. (1979). An analytical method of identifying biased test items. *Journal of Experimental Education, 48,* 153–154.

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

Wright, B. J. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Authors

GREGORY CAMILLI, Lecturer, Rutgers University, Graduate School of Education, 10 Seminary Place, New Brunswick, NJ 08903. *Specializations:* program evaluation, applied statistics.

LORRIE A. SHEPARD, Professor, University of Colorado, School of Education, Campus Box 249, Boulder, CO 80309. *Specializations:* measurement and policy research.