A New Approach to Assessing Test Bias

Adam W. Meade
North Carolina State University

Michael Fetzer
PreVisor

A new regression-based method of assessing test bias is proposed. We highlight two different potential causes of differences in groups’ regression line intercepts. Intercepts differing due to mean criterion score differences are not interpreted as predictive test bias. Using both simulated and employee data, we illustrate this new approach.

Test bias is a systematic error in how a test measures members of a particular group (Camilli & Shepard, 1994). Test bias is a fundamentally important issue in testing as pervasive and systematic sources of error can lead to erroneous inferences regarding the interpretation and use of test scores. Test bias may preclude comparisons across groups and is cause for considerable concern in decision making contexts such as employee selection (U.S. EEOC, 1978).

The most common method of assessing test bias is the Cleary (1968) method, which involves testing for both slope and intercept regression line differences across groups (Sackett & Wilk, 1994). In this approach, the test serves as the predictor variable and some measure of performance serves as the criterion. The Cleary approach is widely used and endorsed by both the Standards for Educational and Psychological Testing (American Educational Research Association/American Psychological Association/National Council on Measurement in Education, 1999) and the Principles for the Validation and Use of Personnel Selection Procedures (Society of Industrial and Organizational Psychology, 2003).

We briefly review the Cleary (1968) approach for assessing test bias, delineate two fundamentally different causes of group intercept differences, discuss potential reasons for these differences, and recommend a new approach to assessing test bias that allows for the distinction of intercept differences that can be said to be due to test bias and those that may be due to other causes. Significantly, we propose a method by which differences in group intercepts may not be interpreted as test bias.

The Cleary (1968) Approach

The Cleary (1968) approach to assessing test bias involves regressing a criterion such as performance on a predictor (i.e., the test), a dichotomous group membership variable, and the cross-product of the predictor and group membership in a single regression model (Sackett, Laczo, & Lippe, 2003). Using this framework, if the group membership variable is significant, the groups are said to have regression lines with different intercepts. If the group-predictor cross-product is significant, groups have regression lines with different slopes. We only examine differences in groups’ regression line intercepts. Like others (e.g., Sackett & Wilk, 1994), we agree that differential slopes indicate differential test validity which is problematic for any test.

While the Cleary (1968) approach to assessing test bias has been widely used with ample success (Schmidt, 1988; Wigdor & Garner, 1982), we contend that there are two fundamentally different potential sources of differences in regression intercepts. The ovals in Figures 1 and 2 represent the scatter plots of examinee test and criterion scores around the Group 1 and Group 2 regression lines. In both, the regression lines have differing intercepts (and equal slopes) for two groups. However, in Figure 1 the two groups differ in their mean test scores but do not differ on the performance criterion. The converse is true in Figure 2; the cause of the differing intercepts is due to mean differences in performance (with no corresponding differences in test scores). Different regression line intercepts for groups will occur whenever there are (a) group mean differences on the criterion that are not proportional to corresponding differences on the predictor, (b)
group mean differences on the predictor that are not proportional to corresponding differences on the criterion, or (c) group mean differences on both the predictor and criterion that are not proportional to one another. We posit that the nature of the cause of differing group intercepts should have some bearing on the interpretation of this difference.

If there are no significant differences in performance (Y1=Y2), the expectation should be that there are no differences in test scores (X1=X2). In light of the evidence of equal standing on performance, one would expect corresponding equal standing on the test. If there are group differences in test scores with no, or disproportionately smaller, differences in criterion scores (see Figure 1), a difference in regression line intercepts will occur that can accurately be interpreted as test bias.

Correspondingly, if there are group differences in the performance criterion (Y1≠Y2), one would expect corresponding proportional differences in test scores (X1≠X2). This is consistent with the notion that a single regression line will fit for multiple demographic groups when test bias is not present. If both the test and performance scores differ in a consistent proportion across groups, the same regression line will fit and the test will tend to show adverse impact, but not test bias (see Figure 3).

However, if there are differences in performance, but no differences in test scores (Figure 2), this is not necessarily due to test bias. Test bias is defined a systematic source of error variance in test scores (Camilli & Shepard, 1994), not criterion scores. If a test shows no mean differences across groups, yet there are mean differences in criterion performance scores, equating this to the fault of the test (and not the criterion) places an overly strong assumption on the quality of the criterion scores and the lack of quality of test scores. This seems especially unusual given that the psychometric properties of tests are routinely considered to be of higher quality than those of criteria in many contexts (Schmitt & Landy, 1993). For example, discussions of the “criterion problem” in job performance measurement have been routine for over 80 years (e.g., Bennett, Lance, & Woehr, 2006; Bingham, 1926) with no clear resolution in sight (Guion, 1998; Thayer, 1992).

Potential Causes for Criterion Mean Differences

There are three primary sources of criterion mean differences: (a) groups could truly differ on the criterion (i.e., a true score difference), (b) there could be systematic bias in the criterion measure, or (c) the difference is due to random error associated with unreliability in the criterion measure.

If true score differences exist on the criterion, but not on the predictor, it could be that an unmeasured variable explains these differences. Sackett et al. (2003) recently illustrated the potential confounding effects of omitted variables on the Cleary (1968) procedure. Using simulated and U.S. Army data, they used personality and cognitive ability predictors in which (a) both were related to performance, and (b) there were group differences on ability but not personality. A regression equation including personality but not ability showed differences in the group intercepts because cognitive ability explains true score differences in performance, but when omitted, the regression of performance on the personality test results in differing intercepts. Thus, it would appear that different regression line intercepts may be due to true score differences in performance that could be accounted for by an omitted variable (James, 1980; Mauro, 1990). Such variables may include ability determinants of performance, such as cognitive ability, but also any of a number of variables that may determine performance, such as differences in environmental or biographical variables, education, prior experience, etc.

A second potential source of criterion mean differences is systematic bias in performance ratings. Several authors have posited that there is minimal racial bias in performance ratings (Landy, Shankster, & Kohler, 1994; Latham & Wexley, 1994; Waldman & Avolio, 1991). However, a recent re-analysis of a large scale field study (Sackett & DuBois, 1991) found that when the same employees were rated by both Black and White supervisors, White employees were given roughly the same rating. However, Black employees were rated roughly 3 of a standard deviation lower by White supervisors than they were Black supervisors (Stauffer & Buckley, 2005). That is to say, the race of the rater mattered little for White employees, but mattered a good deal for Black employees. Stauffer and Buckley (2005) go as far as to say that the evidence of racial bias in performance ratings is clear.

Similarly, gender bias in performance ratings has been shown for stereotypical male jobs in lab studies (Swim, Borgida, Maruyama, & Myers, 1989) and in field studies (Eagly, Makhijani, & Klonsky, 1992; Pazy & Oron, 2001). Additionally, a recent study of performance ratings among high-level managers found that women received lower performance ratings for male stereotyped jobs than for female stereotyped jobs, while no such effects were found for male employees. The authors concluded that gender bias was present in performance ratings under some conditions (Lyness & Heilman, 2006).

Lastly, several authors have bemoaned the
lack of diligence often encountered in performance measurement (Guion, 1998; Thayer, 1992). With additional random error comes additional opportunity to observe differences in group means due to random error, which may in turn lead to significant differences in group regression line intercepts.

In sum, the Cleary (1968) test of differences in regression lines is threatened in two ways: (a) true score differences in performance may not be fully accounted for by the measured predictors (but may be accounted for by unmeasured variables), thus regression intercepts will differ via no fault of the test, and (b) there could be bias or random error present in performance ratings themselves which results in different intercepts even when no test bias is present.

A New Procedure

We suggest a new process for evaluating test bias in which the cause of differences in intercepts is considered in the interpretation of results.

Step 1: Conduct a t-test for differences in the predictors across groups. If significant, compute an d effect size estimate for differences in groups’ predictors scores.

Step 2: Regress performance on group membership to test for mean differences on the criterion. If significant differences are found, compute an d effect size for group differences on the criterion.

Step 3: Regress performance on group membership, test scores, and the interaction between the two (i.e., the Cleary, 1968, regression model). A significant group membership regression coefficient indicates differing regression line intercepts for the groups while a significant group-predictor interaction indicates a different regression line slope for the groups.

Table 1 provides guidance on how to interpret results. Differences in slopes are indicative of bias and are not depicted in Table 1. Additionally, a lack of differences in intercepts indicates a lack of bias. To this extent, the current approach matches the Clearly (1968) approach.

The proposed procedure differs from the Cleary approach, however, in the two italicized cells of Table 1. In the current procedure, a difference in intercepts is typically not indicative of bias when there are no significant differences in groups’ predictor scores. If there are differences in performance but not test scores, it is more likely that the test is not biased but that other explanatory factors that are omitted from the regression model explain group differences in performance, or that the criterion is biased or flawed in some way. Conversely, if performance scores do not differ across groups but test scores do differ across groups, the difference in regression line intercepts is likely due to a problem with the test (i.e., test bias).

If there are significant differences in intercepts, and there are differences in both predictor and criterion scores, it is up to the researcher to use his or her judgment as to the nature of these differences. D effect size estimates for mean differences on the predictor and criterion can shed light on the situation. For example, if the difference in test scores is quite large (e.g., d = 1.0) while the difference in criterion scores is disproportionately smaller (e.g., d = .2) then the test should be considered biased in absence of information from other predictors. The d statistic will be particularly useful for interpretation with large sample sizes which will result in high power for the statistical tests used to evaluate bias. We now illustrate this method with both simulated and job incumbent data.

Simulation Study

Method

Two conditions of simulation data were created, each with 1000 simulated respondents, 800 of which were simulated to be majority group members and 200 minority group members. In the first, we replicated the simulation method of Sackett et al. (2003). Specifically, we created responses to a conscientiousness measure and a cognitive ability test that were normally distributed. We also simulated random error and made performance equal to the sum of cognitive ability, conscientiousness, and two times the error component (cf. Sackett et al., 2003). This procedure results in predictors that account for approximately half of the variance in performance.

In Condition 1, like Sackett et al. (2003), we simulated population group mean differences of one standard deviation on the cognitive ability test score. Though not noted in Sackett et al., this has the effect of also creating a corresponding group mean difference of one standard unit in performance scores favoring the majority group as performance is linearly determined by ability (i.e., a true score difference in performance explained by a true score difference in cognitive ability). This condition represents a case of no bias in the predictors, but mean differences in both performance and cognitive ability. Thus, Condition 1 corresponds directly to Figure 1 with respect to the personality predictor and to Figure 3 with respect to cognitive ability.
In Condition 2, we held performance constant across groups by setting the performance score equal to the same linear composite as Condition 1 but also adding a constant of 1.0 to the performance scores of the minority group to offset the constant added to the majority group performance scores (via the cognitive ability predictor). Condition 2 represents the situation depicted in Figure 2 with respect to cognitive ability; there is no bias for the personality predictor with no mean differences on personality or performance.

Analyses. We followed the proposed outline of analyses described in the three steps earlier. Separate regressions were conducted for each predictor as would be the case if only one of the two predictors had been administered to the applicant sample (cf. Sackett et al., 2003).

Results

Condition 1. As expected, there were no significant differences on personality test scores, t(998)=.68, p>.05, yet there were significant differences on cognitive ability test scores t(998)=14.23, p<.001, d=1.13. The initial regression of performance on the group membership variable revealed that there were differences in group performance, F(1,998)=27.69, p<.001, b=.99, d=.42.

In the regression of performance onto personality, group membership, and the interaction between the two, as expected, both the personality, t(1)=5.23, p<.001, b=.83, and group membership, t(1)=5.45, p<.001, b=.94, predictors were significant, while the interaction term was not. See Figure 4 for the regression lines obtained from these analyses.

In the regression of performance on cognitive ability, group membership, and the interaction between the two, as expected the cognitive ability predictor was significant, t(1)=5.81, p<.001, b=.87, while the other two predictors were not.

Condition 2. As expected, there were no significant differences on personality test scores, t=.51, p>.05, yet were significant differences on cognitive ability test scores, t=12.40, p<.001, d=.98. There were no differences in group performance, F(1,998)=.47, p>.05.

In the regression of performance onto conscientiousness, group membership, and the interaction between the two, only the conscientiousness variable was significant, t(1)=6.97, p<.001, b=1.09. Figure 4 also depicts these regression lines.

In the regression of performance on cognitive ability, group membership, and the interaction between the two, the cognitive ability predictor was significant, t(1)=7.31, p<.001, b=1.16, as was the group membership predictor, t(1)=5.72, p<.001, b=1.41. Given no differences in performance and differing intercepts, we can infer bias in the cognitive ability predictor.

Discussion

In Condition 2, both the Cleary method and the method proposed here result in identical conclusions with both procedures accurately detecting bias in cognitive ability and both procedures accurately determining no bias for conscientiousness. In Condition 1, both procedures also accurately found no bias for cognitive ability. However, in Condition 1, we know that the omitted cognitive ability variable is responsible for the true score differences in performance and that conscientiousness is not biased. Importantly, the Cleary (1968) procedure would result in the conclusion that the conscientiousness variable is biased across groups while the current procedure results in the determination that there is no bias.

Job Incumbent Sample

Method

Data consisted of 3,282 employees in clerical-related jobs (e.g., tellers, bookkeepers) at several financial institutions. Predictors included a biodata measure of Problem Solving and Flexibility, which is a sub-scale of the Customer Service and Clerical Potential Index (Erwin, 2001a) and the Test of Learning Ability (Erwin, 2001b). The learning ability test is a timed 108 item cognitive ability test measuring vocabulary, arithmetic reasoning, and spatial relation processing.

The criterion was composed of averages across two sets of ratings (immediate supervisors and one additional source) averaged across 28 job task performance items (e.g., “Pays attention to detail”) that were derived from a detailed job analysis. Factor analysis indicated that a single underlying “task performance” factor best fit the data.

Analyses

Given that group differences in cognitive ability tests are often most pronounced for Black-White racial group comparisons (Neisser et al., 1996), we focused our illustration on these two groups. We followed the same procedure as before, except that an additional regression model simultaneously examined both biodata and learning ability as well as their cross-product.

Results

The t-test for group differences in biodata scores was not significant, t(3280)=0.69, p=.49,
d=.03, while that of the Learning Ability test was significant, t(2773)=27.07, p<.001, d=1.13. Descriptive statistics are available in Table 2.

Results from the first regression indicated that the two racial groups differ in terms of their performance ratings (see Table 3; d=.37). In the second model, the biodata variable was significant, as was the race variable indicating different group intercepts. The interaction term was not significant, indicating equal slopes. The differences observed in intercepts of the biodata regression lines, with a lack of differences in mean biodata scores across groups, would be indicative of the situation depicted in Figure 2 and is not considered bias.

The third model illustrates the effect of entering the cognitive ability predictor into the model, which resulted in the racial group main effect variable being non-significant (see Table 3). In other words, once the regression fully accounts for learning ability, there is no indicated racial bias in the predictors.

General Discussion

In this paper, we propose a new approach for evaluating test bias that retains the emphasis on equivalent group regression lines. As with the Cleary (1968) approach, equivalent regression lines are indicative of a lack of bias and differing slopes across groups indicates bias in the test. However, the Cleary approach suggests that differences in regression line intercepts across groups should uniformly be interpreted as test bias whereas the proposed approach differentiates between intercept differences due to group mean differences on the predictor and those on the criterion. Both approaches treat differences in intercepts as bias when groups do not differ on the criterion but differ on their test scores. Unlike the Clearly approach, however, group differences on the criterion but not on test scores is not treated as bias. More likely, there is a factor other than test scores that account for differences on the criterion variable. Such sources include true group mean differences on the criterion that may be accounted for by an omitted variable, bias in the criterion ratings, or random error associated with poor criterion reliability. The Cleary approach treats this situation as a problem with the test whereas the proposed alternative method treats differences in criterion scores as a problem with criterion measurement, or lack of accounting for true score differences in the criterion.

Limitations and Recommendations

One limitation is that there is some need for judgment in the current method. Specifically, when intercepts differ across groups and both performance and predictor scores differ as well, it is difficult to determine whether the differing intercepts are due to criterion problems, test bias, or both. We recommend evaluating the magnitude of group differences using effect size estimates, though we do not provide firm rules as to the interpretation of these. When in doubt, however, we would encourage researchers to presume differing intercepts are due to test bias when it is not clear that criterion differences are the source of the differing intercepts. A second limitation of the proposed method is that like the Cleary (1968) method, this procedure is dependent upon the use of significance tests, the power of which is highly dependent upon sample size.

There were limitations to the simulation and applied demonstrations as well. As with any simulation study, simulated data are idealized and externally valid only to the extent to which they are reflective of the true state of affairs. The applied sample data used to illustrate this method was also not ideal. In general, cognitive predictors tend to account for more criterion variance than was the case here (Schmidt & Hunter, 1998).

In sum, we believe that the proposed method for evaluating test bias has great potential to aid researchers in differentiating between different sources of group differences in the selection context. We encourage researchers and test developers to evaluate test bias by first examining group mean differences in both predictors and criteria before proceeding to conventional tests of group differences in regression line intercepts and slopes.

References


Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated


**New Approach to Test Bias**

[References not shown]
### Table 1.

**Interpretation of Tests of Bias**

<table>
<thead>
<tr>
<th>Intercept Test</th>
<th>Predictor Test</th>
<th>Criterion Test</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not significant</td>
<td>Not Significant</td>
<td>Not Significant</td>
<td>No bias</td>
</tr>
<tr>
<td></td>
<td>Significant</td>
<td>Not Significant</td>
<td>No bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Significant</td>
<td>No bias</td>
</tr>
<tr>
<td>Significant</td>
<td>Not Significant</td>
<td>Not Significant</td>
<td>Bias</td>
</tr>
<tr>
<td></td>
<td>Significant</td>
<td>Not Significant</td>
<td><strong>No Bias</strong></td>
</tr>
<tr>
<td>Significant</td>
<td>Significant</td>
<td>Not Significant</td>
<td>Bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Significant</td>
<td><strong>Maybe Bias</strong></td>
</tr>
</tbody>
</table>

*Note: Italicized conclusions indicate different interpretations than those under the Cleary (1968) procedure.*
Table 2.

Descriptive Statistics for Variables

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th></th>
<th>Black</th>
<th></th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Performance Rating</td>
<td>5.50</td>
<td>1.08</td>
<td>5.09</td>
<td>1.12</td>
<td>.37</td>
</tr>
<tr>
<td>Biodata</td>
<td>8.64</td>
<td>2.10</td>
<td>8.58</td>
<td>2.06</td>
<td>.03</td>
</tr>
<tr>
<td>Learning Ability</td>
<td>70.83</td>
<td>12.49</td>
<td>56.64</td>
<td>12.60</td>
<td>1.13</td>
</tr>
</tbody>
</table>

Note: White N=2323, Black N = 959 for Performance Ratings and Biodata. White N=1973, Black N=802 for Learning Ability.

Table 3.

Regression Coefficients from Hierarchical Regression

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictor</th>
<th>b</th>
<th>SE</th>
<th>β</th>
<th>t</th>
<th>p</th>
<th>R²</th>
<th>(ΔR²)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Race</td>
<td>0.38</td>
<td>0.05</td>
<td>0.16</td>
<td>8.41</td>
<td>0.00</td>
<td>.025</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Race</td>
<td>0.38</td>
<td>0.04</td>
<td>0.16</td>
<td>8.62</td>
<td>0.00</td>
<td>.071</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biodata</td>
<td>0.20</td>
<td>0.04</td>
<td>0.18</td>
<td>5.30</td>
<td>0.00</td>
<td>(.046)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Race*Biodata</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.98</td>
<td>0.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Race</td>
<td>-0.38</td>
<td>0.22</td>
<td>-0.16</td>
<td>-1.74</td>
<td>0.08</td>
<td>.111</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biodata</td>
<td>0.18</td>
<td>0.04</td>
<td>0.16</td>
<td>4.70</td>
<td>0.00</td>
<td>(.040)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Race*Biodata</td>
<td>0.01</td>
<td>0.04</td>
<td>0.01</td>
<td>0.20</td>
<td>0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Learning Ability (LA)</td>
<td>0.04</td>
<td>0.01</td>
<td>0.15</td>
<td>3.88</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Race*LA</td>
<td>0.03</td>
<td>0.01</td>
<td>0.26</td>
<td>2.41</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N=2764
Figure 1. Different regression intercepts due to mean differences on predictor.

![Mean Difference on Predictor](image1)

Figure 2. Different regression intercepts due to mean differences on criterion.

![Mean Difference on Criterion](image2)

Figure 3. Equal intercepts with mean differences on predictor and criterion.

![Mean Difference on Predictor and Criterion - No Bias](image3)
Figure 4. Regression lines from simulation study.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Cognitive Ability</th>
<th>Conscientiousness</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Majority</td>
<td>Majority</td>
</tr>
<tr>
<td></td>
<td>Minority</td>
<td>Minority</td>
</tr>
<tr>
<td>B</td>
<td>Majority</td>
<td>Majority</td>
</tr>
<tr>
<td></td>
<td>Minority</td>
<td>Minority</td>
</tr>
</tbody>
</table>