Industrial and Organizational Psychology, **3** (2010), 213–217.

COMMENTARIES

Conceptual and Technical Issues in Conducting and Interpreting Differential Prediction Analyses

PAUL R. SACKETT
University of Minnesota

PHILIP BOBKO
Gettysburg College

We briefly review the focal article by Meade and Tonidandel (2010) regarding the Cleary approach for investigating differences in subgroup regression lines. Their efforts may be motivationally helpful to the literature, yet we believe that some of their recommended procedures and specific conclusions are flawed. Overall, we suggest that

- Meade and Tonidandel's concerns are overstated, and they do not sufficiently define the context of their analysis;
- there is an imbalance in the critical scrutiny given to internal and external methods of assessing test bias and, in turn, the feasibility of their suggestions about differential item functioning (DIF) is questionable;
- there are ambiguities in inferences they draw from specific data patterns; and
- there is an important statistical flaw in their proposed regression approach.

Correspondence concerning this article should be addressed to Paul R. Sackett.
E-mail: psackett@umn.edu
Address: Department of Psychology, University of Minnesota, Elliott Hall, 75 East River Road, Minneapolis, MN 55455
Paul R. Sackett, University of Minnesota; Philip Bobko, Departments of Management and Psychology, Gettysburg College.

We offer an initial set of observations about these various aspects of their suggestions (while noting there are other issues in Meade and Tonidandel's work that require additional analyses in our future efforts).

Overstating the Problem

Like many fields of endeavor, understanding of the concept of "test bias" has evolved over time. Meade and Tonidandel, as well as Meade and Fetzer (2009), cite early studies as the basis for claiming misunderstanding and confusion. But the most recent and relevant set of professional standards, namely, the Society for Industrial and Organizational Psychology (SIOP) *Principles* (2003), raise many of the concerns about differential prediction raised in their article, including differentiating between predictive bias and measurement bias, the need for an unbiased criterion, and concern over low statistical power (including the role of predictor unreliability and range restriction). The *Principles* do not address the issue of omitted variables, which has gained prominence since the *Principles* were completed (e.g., Sackett, Laczo, & Lippe, 2003). We suggest that attention to the omitted variables problem can be useful. But, the *Principles* show that the field has indeed recognized the complexities of differential

https://doi.org/10.1111/j.1754-9434.2010.01226.x Published online by Cambridge University Press

prediction, including many of the central issues raised by Meade and Tonidandel.

Overall, we also suspect that "test bias" is used as shorthand in the same way that academics and practitioners use "test validity." When pressed, we believe that many in our field would say "yes, I know that tests per se are not valid or invalid; the question of interest is whether inferences drawn from tests in particular types of settings for particular types of purposes are valid." The technically precise language is a mouthful, and "test validity" is understood to stand in for the complete phrase. Similarly, it is not clear that those using the term "test bias" today fail to understand that predictive bias refers to bias in the use of a set of test scores to predict a given criterion, rather than to an inherent problem with a test, per se. That said, we agree it is useful to consider "predictive bias" and "measurement bias," or equivalent terms, as separable constructs.

Need for Clarity as to the Purpose of Differential Prediction Analysis

Differential prediction analyses can be conducted for at least two somewhat distinct purposes: to comply with the regulatory framework under which selection practice often operates or to provide the scientist/practitioner with insight as to the nature of predictor/criterion relationships in the setting of interest. Although not explicitly stated by Meade and Tonidandel, their paper appears to be focusing on the second purpose. For the first, a finding of group differences often constitutes the point of entry to regulatory scrutiny; it is in the context of a finding of adverse impact that an investigation of differential prediction comes into play (cf. *Uniform Guidelines*, U.S. Equal Employment Opportunity Commission, 1978, Section 1607.14B8[b]). Meade and Tonidandel state that they prefer to examine differential prediction regardless of the presence or absence of mean differences on the predictor, which takes their inquiry well beyond that of a regulatory framework.

Imbalance in the Critical Scrutiny Given to Internal and External Methods

As noted earlier and in the literature, there are internal ("measurement bias") and external ("predictive bias") approaches to test bias. These conceptual phrases address different things. Measurement bias focuses on the question of whether individuals with the same standing on the construct of interest have the same expected test (or item) score regardless of group membership; predictive bias focuses on whether a given test score leads to the same predicted criterion score regardless of group membership. Meade and Tonidandel put predictive bias under close scrutiny and document a range of complexities in using and interpreting the results of predictive bias analyses that have been identified in the literature over the years. Given these complexities, they strongly advocate the use of internal methods, noting that "very often" such an analysis might answer "the question that researchers would most like to answer." However, given that measurement bias and differential prediction (or predictive bias) are conceptually separable, as they and others note, then internal analyses do not answer "the question," but they answer "a question" (e.g., internal analyses do not directly address predictive bias in operational test use). So, there could be a logical fallacy in their thinking in that their strong recommendation about internal analyses does not necessarily follow from concerns with predictive bias methodologies (and/or the questions those latter methods help answer).

In addition, Meade and Tonidandel do not subject internal methods to any critical scrutiny, much less the close scrutiny they apply to external methods. This short commentary is not the place for a detailed examination of internal methods, but a few points are worth raising. First, they appear to be thinking narrowly about one type of predictor, namely, multi-item tests. It is in that context that the possibility of examining individual test items for DIF arises. The

https://doi.org/10.1111/j.1754-9434.2010.01226.x Published online by Cambridge University Press

personnel selection field sometimes uses a range of predictors that may produce a single score, such as an overall rating in an interview or an assessment center. While the use of such measures to predict a given criterion can be examined for predictive bias, DIF analysis is not always applicable. Second, consider the discussion of internal methods in the SIOP *Principles* (2003). The *Principles* note a range of difficulties in applying DIF analysis, including the need for data on large research samples prior to operational use, the requirement of unidimensional tests when many widely used predictors are potentially heterogenous (e.g., situational judgment tests, biodata inventories), and the common finding of approximately equal numbers of items favoring each group, which results in no systematic bias at the test level (see also Hunter & Schmidt, 2000 for another discussion of similar concerns). In fact, the *Principles* conclude that DIF findings should be viewed with caution and that "DIF analysis is not likely to become a routine or expected part of the test development and validation process in employment settings" (p. 34).

Thus, although there may be instances where internal methods are useful, we disagree with Meade and Tonidandel's recommendation that internal methods always be used and with the implication that failure to use these methods constitutes inappropriate behavior on the part of a selection researcher/practitioner.

The Potential Error of Interpreting Main Effects in the Presence of Interactions

Meade and Tonidandel briefly summarize analyses presented by Meade and Fetzer (2009) related to the examination of the omitted variables problem. In a nutshell, they report an analysis of differential prediction by race in the use of a biodata scale and indicate that a significant race effect becomes nonsignificant when an ability test is added to the model. Conceptually, this would be an illustration of the omitted variable problem; that is, in the model when ability is omitted, the variance shared between ability and race is attributed to race.

However, Meade and Tonidandel do not accurately describe the Meade and Fetzer (2009) study. Not only did Meade and Fetzer add ability to the model but they also added a Race × Ability interaction term.[1] Furthermore, unless the variables are at the ratio level, regression weights for main effects are arbitrary and problematic in the presence of any nonzero interaction terms in the model (e.g., Bobko, 2001; Cohen & Cohen, 1983; Gocka, 1974). In interactive models, the scaling choices for either of the main effect variables can influence the regression weight of the other main effect variable. More specifically, the scaling of the group membership variable is arbitrary (the coding could be 0, 1 or 1, 0 or −1, +1, etc.), as is the scaling of the test variable (*z*-scores, 0–10, 70–100, etc.). In turn, the resultant main effects (i.e., the race regression weights being misinterpreted by Meade and Fetzer in their procedure) are arbitrary. With a nonzero interaction in the model, values of the main effect regression weights could be made smaller or larger, become zero, or even change sign by using a different coding scheme for the subgrouping or test score variable.[2] Thus, Meade and

1. Meade and Tonidandel also appear to claim that Meade and Fetzer found that "a common regression line fit the data" when a cognitive ability test was added. This is incorrect, as the interaction term was statistically significant.
2. This statement is true whether or not the nonzero interaction term is statistically significant and different from zero. Although beyond the scope of this brief note, one might develop a procedure that is contingent upon the statistical significance of the interaction weight. Also, we thank one reviewer for pointing out that, if the variables are centered, the main effect weight is an average of the interactive differences. However, the average of the interactive regression effect does not necessarily have much meaning in a selection context. For example, in disordinal interactions, the average contains both positive and negative effects. And, in ordinal interactions, the average contains a variety of possibly quite different magnitudes of the interactive effects. Within a selection context, this might not be helpful, as the operational magnitude of any interactive effects might depend substantially on the cut-point used in selection.

https://doi.org/10.1111/j.1754-9434.2010.01226.x Published online by Cambridge University Press

Fetzer's procedure (and implicitly Meade and Tonidandel's procedure) is statistically flawed and should not be used as recommended.

Ambiguities in Inferences Drawn From Specific Data Patterns

We endorse the notion that it can be helpful to consider contextual features of the setting in which a differential prediction analysis is conducted. The more the details of the setting are known and considered, the greater the likelihood that plausible hypotheses about the features that drive a finding of differential prediction can be formed and subsequently addressed. In specific settings, some explanations can be ruled out on logical grounds (e.g., the possibility of rater bias in the criterion might be ruled out in a setting where the criterion is an objectively scored post-training knowledge test), and other explanations can be identified as potentially testable drivers of the finding of predictive bias. But, we are not persuaded by Meade and Tonidandel's analyses of various generic situations.

For example, their Scenario 1 involves mean differences on the test but no differences on the criterion. They attribute to Meade and Fetzer (2009) the conclusion that such a finding is "most likely" due to measurement bias in the test, and they conclude that "it is difficult to imagine any scenario under which use of the test in this setting would be considered fair."

On the contrary, one can readily imagine settings where this pattern would be produced, where the differential prediction is not due to bias in the test, and where test use within a selection system would indeed be considered fair. In particular, this could occur in settings where the criterion is influenced by two predictor constructs, with comparable group differences in opposite directions on the two predictors. For example, men tend to score higher on math tests than women, whereas women tend to score higher on verbal tests than men (Sackett, Borneman, & Connelly, 2008). If the criterion is equally influenced by math and verbal ability, then Meade and Tonidandel's Scenario 1 would be expected if the math test alone is examined for differential prediction; namely, a mean gender difference on the test and none on the criterion. Thus, contrary to the suggestion that such a finding is "most likely" due to test bias, the issue could readily be another instance of an omitted variables problem.

Continuing the example further, if the selection system were to involve both a math and a verbal test, it is the composite of math and verbal that would be assessed for differential prediction. The expected finding would be no mean difference on the composite, and the result is a scenario of no mean difference on either the predictor or the criterion. Elsewhere in the paper Meade and Tonidandel do note that differential prediction analysis should be applied to a composite of all predictors in a selection system to avoid an omitted variables problem. However, they fail to mention this possibility when they apply differential prediction analyses to single tests in their scenarios.

Conclusion

We appreciate the notion of probing into the bases for findings of differential prediction. We suggest, though, that such an inquiry is best accomplished in the particular context of the selection system in question, rather than by following generic guidelines such as "if there are predictor differences but no criterion differences it's most likely due to measurement bias in the test." We also advise careful attention to the statistical procedures used for differential prediction analyses, with particular eye to the issue of misinterpreting main effects in the presence of interactions. We also suggest that internal methods of assessing measurement bias merit the same careful scrutiny as external methods of assessing predictive bias and caution against a call for universally strong advocacy of internal methods. We appreciate the opportunity to comment on their paper and look forward to yet further clarity in this domain.

https://doi.org/10.1111/j.1754-9434.2010.01226.x Published online by Cambridge University Press

References

Bobko, P. (2001). *Correlation and regression: Principles and applications for industrial/organizational psychology and management* (2nd ed.). Thousand Oaks, CA: Sage.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Gocka, E. (1974). Coding for correlation and regression. *Educational and Psychological Measurement, 34,* 771–783.

Hunter, J., & Schmidt, F. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6,* 151–158.

Meade, A. W., & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods, 12,* 738–761.

Meade, A. W., & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology, 3,* 192–205.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63,* 215–227.

Sackett, P. R., Laczo, R. M., and Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology, 88,* 1046–1056.

Society for Industrial and Organizational Psychology (SIOP). (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Society for Industrial and Organizational Psychology.

U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, U.S. Department of Justice (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38295–38309.

https://doi.org/10.1111/j.1754-9434.2010.01226.x Published online by Cambridge University Press