Industrial and Organizational Psychology, **3** (2010), 206–209.

COMMENTARIES

The Need for Even Further Clarity About Cleary

STEVEN F. CRONSHAW
University of Northern British Columbia

GREG A. CHUNG-YAN
University of Windsor

In addition to providing a very good primer on conventional thinking on test bias as assessed by the Cleary model, Meade and Tonidandel (2010) present some provocative recommendations for future work in this area. Most test developers and users will agree with their second recommendation—to always examine both differential functioning and prediction (when possible). Many will agree with their third recommendation that tests can sometimes be used even when differential prediction is found. Their first recommendation—that the field stop using the term test bias—will meet with widespread disagreement and even consternation. In this commentary, we agree with Meade and Tonidandel's first recommendation, although not necessarily for the same reasons that they give, but go further to make additional recommendations that build on their very good start in critiquing the state of testing research and practice.

Meade and Tonidandel's concerns about the imprecision and ambiguity in the term "test bias" is well placed but is only the tip of a larger problem that remains submerged below the waterline of the psychological testing literature. Inordinately broad and imprecise nouns often become associated with multiple statistical formulations, each of which is precise but different from, and sometimes inconsistent with, cognate others. The noun compound "test bias" is an excellent example of this problem: The widely cited book on mental testing by Jensen (1980) describes a number of different mutually exclusive sets of statistical criteria for defining test bias, the Cleary (1968) model being but one. Although the conventional definition of test bias subsumes only two types of bias (measurement bias and differential prediction) as pointed out by Meade and Tonidandel, numerous specific statistical means also have been proposed for determining whether test bias exists (e.g., the Thorndike, 1971, model). As a result, "test bias" has different denotations as well as connotations for different people, to the point where the term has ceased to have any meaning beyond a broad label applied, most often pejoratively, to tests and measures in educational and psychological testing. The reaction to such broad and imprecise, but emotionally loaded, labeling is predictable enough. Because a single label has the power to brand a test or class of tests as defective at best or socially pernicious at worst, testing stakeholders with conflicting interests in

Correspondence concerning this article should be addressed to Steven F. Cronshaw.
E-mail: cronshaw@unbc.ca
Address: School of Business, University of Northern British Columbia, Prince George, British Columbia, Canada V2N 4Z9
Steven F. Cronshaw, School of Business, University of Northern British Columbia; Greg A. Chung-Yan, Department of Psychology, University of Windsor.

https://doi.org/10.1111/j.1754-9434.2010.01224.x Published online by Cambridge University Press

psychological testing feel the need to adopt an "all-or-none" mentality with regard to the presence or absence of "test bias." Test developers and psychologists using tests are under considerable pressure to find tests unbiased; antitesting advocates take the position that tests, especially intelligence tests, are always culturally biased against groups protected under equal employment opportunity legislation regardless of findings reported in the testing literature.

The emerging consensus of all-or-none opinion in industrial and organizational (I–O) psychology toward the existence of test bias in ability and achievement tests, including tests of intelligence or general cognitive ability, is "none." For example, Hunter and Schmidt (2000) state with respect to racial and gender comparisons "... research has established that total scores on ability and achievement tests are predictively unbiased ..." (p. 151). They go on to argue that "... there appears to be no evidence that items on currently used tests function differently in different racial and gender groups" (p. 151). However, as Meade and Tonidandel point out, the most commonly used technique for determining test bias is *predictive* using the Cleary (1968) model. We would go further to argue that for most I–O psychologists, the Cleary regression-based model is considered to be the *only* legitimate means of determining the presence or absence of test bias based on the overall test score. This belief appears to be due at least in part to an influential paper by Petersen and Novick (1976) that evaluates different statistical models for culture-fair selection. The gist of their argument is expressed in a premise, then a conclusion, as follows:

> Premise: If it is true that "optimal prediction and fairness (lack of bias) are ... strictly equivalent" (p. 5), then
>
> Conclusion: It is also true that culture selection models that do not maximize "... expected performance for each individual and hence overall" (p. 5) are "internally contradictory" (p. 24).

As pointed out by Petersen and Novick (1976), among culture selection models that do not maximize expected performance over both selected and nonselected individuals and groups are the constant ratio (Thorndike, 1971), conditional probability (Cole, 1973), and the equal probability (Linn, 1973) models. The above argument is logically valid in that the premise entails the conclusion (Audi, 1999), but the conclusion only holds necessarily true if the premise is accepted. We do not accept Petersen and Novick's premise, as evidenced in Chung-Yan and Cronshaw (2002) where we reject the premise that the fairness/lack of bias is strictly equivalent to optimal prediction using the Cleary model. We argue that it is up to the individual I–O psychologist based on his or her personal values, professional judgment, and ethical standards to accept or reject the premise underlying the Cleary regression-based model. If the I–O psychologist accepts the premise of Petersen and Novick's argument, then it is appropriate (and even necessary) to conclude that the nonoptimizing models of selection bias/fairness based on the overall test score that are proposed in the test bias literature have been "discredited" (a term used by one of the reviewers of this commentary) or are "internally contradictory" (as stated in the conclusion by Petersen & Novick, p. 24). However, as we have shown, it is in no way foregone that their initial premise should be accepted, and, as a result, it is not predetermined that the conclusion to their argument should be accepted on logical, scientific, or any other grounds.

The narrowing of the I–O definition to a single correct definition of predictive test bias, that is, Cleary's, probably results from I–O psychology's pragmatic need to sharpen the operational meaning of "test bias" and to do so in a way that proves our value as "utility maximizers" to the managers who are our clients, but it does so at the expense of begging the question. But perhaps the field's gravitation to a single correct definition was inevitable given the difficulties created by compressing many differing but technically precise statistical

https://doi.org/10.1111/j.1754-9434.2010.01224.x Published online by Cambridge University Press

definitions under the impossibly ambiguous term of *test bias*. Meade and Tonidandel are right in their injunction to stop using the term *test bias*, but their argument does not run deeply or broadly enough into the reasons why the term should be dropped from the field's lexicon.

Why has the Cleary model achieved near hegemony among I–O psychologists who study problems of test bias? We have already alluded to reasons for this state of affairs: It is generally believed that the client is best served, and the I–O psychologist's value to the organization is convincingly demonstrated to the client, by maximizing the utility of the test or measure for predicting a criterion of I–O effectiveness, efficiency, or productivity. This value dominates in the I–O psychologist's self-professed *raison-d'etre* and provides the vital impulse to I–O interventions generally (training, performance appraisal, goal setting, etc.). There is nothing inherently wrong with this underlying value but it should be recognized for what it is: a statement of subjective value, a desired outcome of organizational affairs, the ultimate good in what I–O psychologists do. It is not surprising that the regression-based Cleary model is preferred by the I–O profession as a model of selection bias/fairness, but the reasons for this have nothing to do with science. The ascendance of the Cleary model is a social and political phenomenon that reflects the I–O psychology field's stake in the contribution we make to organizations and institutions, rather than any inherent scientific superiority of that method over the many others in the selection bias and fairness literature.

Meade and Tonidandel (2010) come closer than most I–O psychologists to acknowledging that test bias and fairness considerations are not separable, although they do not go far enough in exploring the implications of that statement. The conventional wisdom in I–O psychology views test bias as separate from fairness considerations, where test bias is a statistical concept of systematic variance and test fairness applies to test use. However, many notables in the testing field have not shared this opinion. For example, Cole (1973) in her classic paper on bias in selection considers test bias to be the "converse" of selection fairness (p. 237). Messick's (1989) unified validity framework provides a deep and thoughtful analysis of the possible meaning of this converse relationship between bias and fairness. He points out that the appropriateness, meaningfulness, and usefulness of the score-based inferences made from test validation depend in part on the social consequences of the testing. Messick states that "... social values and social consequences cannot be ignored in considerations of validity ..." (p. 19). To the extent that test bias models are an extension of the fundamental validation question (and most would agree that they are), and to the extent that the conventional wisdom associates test fairness with social consequences (and most would agree with this statement as well), test bias and fairness are unavoidably and inextricably bound together in common fate with validity and are indeed "converse" phenomena. Interestingly, Meade and Tonidandel draw a distinction between properties of the test and test use, a position that—considered within the context of Messick's unified validity framework—is inconsistent with the coextensiveness of test bias and fairness. Messick's work suggests that some of Meade and Tonidandel's arguments, although provocative and interesting, need further tightening and reconciliation within the broader validity framework.

The criterion is a complicating factor in the study of selection bias. As Meade and Tonidandel point out, bias in the criterion is one reason for findings of predictive test bias, but criterion bias occurs independently of the technical properties of the test itself. They do not acknowledge, however, that criterion bias is a summary label in its own right. Brogden and Taylor (1950) in their influential paper on criterion bias subclassify it into four sources of "imperfection" in criteria: (a) criterion deficiency, (b) criterion

https://doi.org/10.1111/j.1754-9434.2010.01224.x Published online by Cambridge University Press

contamination, (c) criterion scale unit bias, and (d) criterion distortion. In fact, a closer examination of the criterion bias problem shows its label to be as imprecise and ambiguous as that of test bias. If Meade and Tonidandel's recommendations are to be followed to a logically and internally consistent conclusion, we think it is necessary to include a fourth recommendation in the list at the end of their paper: *Stop using the imprecise and ambiguous term* criterion bias. If the source of differential prediction is thought to reside in the criterion, a precise technical label and definition for the criterion "imperfection" (Brogden & Taylor) that applies to the specific case will serve the field better than a generalized and pseudoexplanatory appeal to "criterion bias."

A similar injunction can, and probably should, be offered for use of the term "test fairness," but we think the point is made. Once we move down the road entered by Meade and Tonidandel, evidence of the need for fundamental language reform accumulates rapidly. The sophistication of research methodology and statistical technique in testing and assessment matters far outstrips our ability to concisely and precisely label and classify the phenomena we study. The solution to this problem is not obvious, but any attempt to rectify the problems of lax language use will have to be foundational and far reaching. Biologists have faced similar problems in their attempts to comprehensively map the genetic structure of plants and animals. Ashburner et al. (2000) argue that unification of genetic sequencing work done worldwide requires a common technical language. They propose a genetic ontology, a "structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism" (p. 26). Perhaps an analogous testing ontology would help us to solve the fundamental language conundrums alluded to by Meade and Tonidandel.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology tool for the unification of biology (The Gene Ontology Consortium). *Nature Genetics, 25,* 25–29. Retrieved on March 6, 2010, from https://bioinformatics.cs.vt.edu/~easychair/Ashburner_NatGenet_2000.pdf).

Audi, R. (1999). *The Cambridge dictionary of philosophy* (2nd ed.). New York: Cambridge University Press.

Brogden, H. E., & Taylor, E. K. (1950). The theory and classification of criterion bias. *Educational and Psychological Measurement, 10,* 159–183.

Chung-Yan, G. A., & Cronshaw, S. F. (2002). A critical re-examination and analysis of cognitive ability tests using the Thorndike model of fairness. *Journal of Occupational and Organizational Psychology, 75,* 489–509.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 237–255.

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

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

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139–161.

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.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). (pp. 13–103). New York: Macmillan.

Petersen, N. S., & Novick, M. R. (1976). Evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3–29.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8,* 63–70.

https://doi.org/10.1111/j.1754-9434.2010.01224.x Published online by Cambridge University Press