DOI: 10.1037//0003-066X.57.2.140

Amplifying Issues Related to Psychological Testing and Assessment

Gregory J. Meyer
University of Alaska Anchorage

Stephen E. Finn
Center for Therapeutic Assessment

Lorraine D. Eyde
U.S. Office of Personnel Management

Gary G. Kay
Georgetown University Medical Center

Robert R. Dies
New Port Richey, FL

Elena J. Eisman
Massachusetts Psychological Association

Tom W. Kubiszyn
University of Texas at Austin

Geoffrey M. Reed
American Psychological Association

We appreciate the comments (Fernández-Ballesteros, 2002, this issue; Garb, Klein, & Grove, 2002, this issue; Hunsley, 2002, this issue; Smith, 2002, this issue) on our article, "Psychological Testing and Psychological Assessment: A Review of Evidence and Issues" (Meyer et al., February 2001). Some points nicely amplify elements of the article, and it is fruitful to briefly consider others.

Garb et al. (2002) noted that many tests were excluded from our review because they have never been meta-analyzed. It is possible this differentially affected inferences about medical tests and perhaps laboratory tests in particular, because 1988 federal regulations mandated proficiency testing to improve their scoring accuracy. However, laboratory test findings are still often erroneous. Up to 20% of the results for common analytes (e.g., cholesterol, glucose) are incorrect in at least three out of five proficiency trials (Hurst, Nickel, & Hilborne, 1998). Also, reports exist of psychological tests with near perfect validity (e.g., r = .98; Ferligoj & Hlebec, 1999), although the analyses target latent constructs and use what we consider to be confounded predictors and criteria (e.g., two self-report scales). Regardless of these issues, our review was based on systematically gathered evidence, and it illustrated the difficulty of differentiating validity for many common medical and psychological tests. Before organizing this evidence, most psychologists, las we did, probably would have anticipated that pulse oximetry, cardiac stress tests, Pap smears, serum cholesterol, dental X-rays, and computed tomography or magnetic resonance imaging scans for various purposes had higher (or perhaps near perfect) validity.

Garb et al. (2002) thought we presented misleading effect sizes (ESs), given the relative risk reduction (RRR) or odds ratios (ORs) reported in the abstracts of two studies. The latter statistics provide useful information, and one can find lively debate about what constitute the proper statistics to compute from the same data. By necessity, we applied a single ES metric to all studies and used the definition that is traditional in psychology, whereby ES magnitude is defined by study size and the statistical significance of the findings (e.g., Rosenthal, 1991). OR and RRR magnitudes cannot be determined by the same information. For a fixed sample size, they may be large and nonsignificant or smaller and significant. Consider the increase from 1 in 1,000 to 3 in 1,000: It is large in a relative sense (200%) but small in an absolute sense (0.2%). Both numbers are informative, but the latter more closely approximates the traditionally defined ES in our review. On a related note, within evidence-based medicine, RRRs and ORs are considered limited (e.g., Sackett, Deeks, & Altman, 1996) because they do not provide the most patient-relevant estimates of effect, which require an absolute rather than a relative scale.

To our task, we clearly brought existing beliefs shaped by our experience that psychological assessments help patients and clinicians with real problems. The evidence we organized spoke to both the strengths and the inherent limitations of test-based information. Simultaneously, the evidence contradicted the assumption that psychological tests are globally inferior to medical tests. In science, evidence should correct such mistaken assumptions, whether they exist among the lay public or among psychologists.

Garb et al. (2002) and Hunsley (2002) suggested that existing evidence does not support multimethod assessments, the limitations of clinical interviews, or the contribution of personality performance tasks to other sources of information. Although we desire more research on these topics, Garb et al.'s and Hunsley's arguments overlooked relevant data and citations in our article, as well as evidence that patient outcomes are improved when treating clinicians receive even minimal test-based feedback (Lambert, Hansen, & Finch, 2001).

Hunsley's (2002) main comments seemed to amplify partial statements of ours in ways we would not. On the basis of those amplifications, it was argued that we were not sufficiently pessimistic or did not address points we tried hard to make. One concern was that we conflated test validity with as-

[illegible]@[illegible]

sessment validity. However, differentiating the two was one of our primary goals. Although both the available evidence and decades of practice-based experience support optimism about documenting the value of well-trained assessment clinicians, a central theme of our article, from abstract to conclusion, was how this has almost never been studied. Thus, we do not think readers would conclude the scientific status of psychological assessment was firmly established, as Hunsley feared, when our recurring point was the opposite. Similarly, after reviewing numerous issues, we suggested "that by relying on a multimethod assessment battery, practitioners have historically used the most efficient means at their disposal to maximize the validity of their judgments about individual clients" (Meyer et al., 2001, p. 150). Although Hunsley agreed that multimethod assessments are beneficial, he criticized a bolder claim that assessment validity must be enhanced no matter what tests are used. We articulated many of the competencies a skilled assessor requires. Understanding distinct methods and the merits of any given scale for assessing a targeted construct is essential. If some have read our article as supporting the haphazard combination of tests, they have seriously misunderstood our position. Fernández-Ballesteros (2002) seemed to articulate a view that meshes with our own, in that choosing the appropriate instruments and constructs for an assessment requires disciplined, evidence-based thinking.

We agree with Smith's (2002) valuable psychometric points (and the questionable utility of tests in his base-rate and counseling center examples), although we believe that they extend our review rather than undermine it. Validity coefficients alone do not tell the whole story about the merits of a test, but they appropriately serve as a central foundation. In the spirit of Smith's comments, we note that most medical and psychological assessments are much more complex than are his examples and require clinicians to continuously update inference probabilities (not fixed base rates) as new and conflicting sources of data are considered. Smith also noted how society may not value psychological processes and raised important questions about utilities. Paralleling Table 2 in our article (Meyer et al., 2001, pp. 136–143), which compared a wide range of validity coefficients, researchers have also compiled tables that compare a wide range of utilities, with utility defined by the cost of a procedure per gain in quality-adjusted life year (QALY). For instance, Graham, Corso, Morris, Segui-Gomez, and Weinstein (1998) reported costs from less than zero per QALY (e.g., 50% use of lap/shoulder belts for auto drivers vs. no restraints) to millions of dollars per QALY (e.g., screening and treating surgeons to prevent HIV transmission, annual vs. biennial Pap smears for women aged 20–75 years). It is unclear where psychological assessments or interventions would fall on such a scale, but we encourage the appropriate research.

Finally, Fernández-Ballesteros (2002) correctly observed that we focused on clinical assessment, which was the charge of our work group, although the issues are often similar for other test-based applications (see Eyde et al., 1993). We look forward to seeing the forthcoming assessment guidelines she described.

REFERENCES

Eyde, L. D., Robertson, G. J., Krug, S. E., Moreland, K. L., Robertson, A. G., Shewan, C. M., Harrison, P. L., Porch, B. E., Hammer, A. L., & Primoff, E. S. (1993). *Responsible test use: Case studies for assessing human behavior*. Washington, DC: American Psychological Association.

Ferligoj, A., & Hlebec, V. (1999). Evaluation of social network measurement instruments. *Social Networks, 21,* 111–130.

Fernández-Ballesteros, R. (2002). How should psychological assessment be considered? *American Psychologist, 57,* 138–139.

Garb, H. N., Klein, D. F., & Grove, W. M. (2002). Comparison of psychological tests and medical tests. *American Psychologist, 57,* 137–138.

Graham, J. D., Corso, P. S., Morris, J. M., Segui-Gomez, M., & Weinstein, M. C. (1998). Evaluating the cost-effectiveness of clinical and public health measures. *Annual Review of Public Health, 19,* 125–152.

Hunsley, J. (2002). Psychological testing and psychological assessment: A closer examination. *American Psychologist, 57,* 139–140.

Hurst, J., Nickel, K., & Hilborne, L. H. (1998). Are physicians' office laboratory results of comparable quality to those produced in other laboratory settings? *JAMA, 279,* 468–471.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69,* 159–172.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

Sackett, D. L., Deeks, J. J., & Altman, D. G. (1996). Down with odds ratios! *Evidence-Based Medicine, 1,* 164–166.

Smith, D. A. (2002). Validity and values: Monetary and otherwise. *American Psychologist, 57,* 136–137.

Correspondence concerning this comment should be addressed to Gregory J. Meyer, Department of Psychology, University of Alaska Anchorage, 3211 Providence Drive, Anchorage, AK 99508. E-mail: afgjm@uaa.alaska.edu