Industrial and Organizational Psychology, **2** (2009), 481–485.
Copyright © 2009 Society for Industrial and Organizational Psychology. 1754-9426/09

"Content" Still Belongs With "Validity"

RICHARD F. TONOWSKI
U.S. Equal Employment Opportunity Commission and University of Maryland University College

Murphy's (2009) main thesis, that the predictive value of a test for job performance is not dependent on the degree to which job content matches test content, is an important statement. But this statement, and others in his article, tends to provoke a "Yes, but . . ." reaction. Here are the main "buts":

- "Job relatedness" is distinguished from "validity" in a way that does not match well with conventional usage in either the psychological or legal spheres.
- There are some situations in which tests are differentiated by content, and content relevance is essential to the testing strategy.
- Content is important for construct validity and for the concept of validity in general. Moreover, the issue for many nonpsychologist stakeholders is that positive manifold or other correlational evidence, rather than content evidence, is neither necessary nor sufficient for validity.

In line with these reservations, the major impact of Murphy's thesis may be not on content validity but on validation transportability and synthetic validity.

Validity and Job Relatedness

Apparently "job relatedness" as discussed by Murphy (p. 453) is meant to correspond to a legal concept, particularly that contained in the Civil Rights Act of 1991. However, Murphy's version does not align well with either the common understandings of job relatedness in industrial and organizational (I–O) psychology or in law. It is something at variance with validity, since a "valid" predictor of performance may not be judged to be "job related."

Murphy is correct in stating that the two terms are not equivalent. From a federal equal employment opportunity (EEO) legal perspective, *validation* is a psychologist's term of art for a way to demonstrate job relatedness. Neither validity nor job relatedness necessarily rests on content matching. EEO law generally is not concerned with job relatedness or validation, unless a facially neutral selection procedure (i.e., the test) has disparate impact on the selection of groups within a protected class. Then there is a very big concern. The Civil Rights Act (1991) defines unlawful practice as "a respondent [employer] uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, gender, or national origin and the respondent fails to demonstrate that the challenged practice is job-related for

Correspondence concerning this article should be addressed to Richard F. Tonowski.
E-mail: richard.tonowski@eeoc.gov
Address: U.S. Equal Employment Opportunity Commission, 131 M Street NE, Room 5NW16H, Washington, DC 20507-0001
Richard Tonowski, U.S. Equal Employment Opportunity Commission and University of Maryland University College, Graduate School of Management and Technology

The views expressed here are those of the author and do not necessarily reflect the views of any agency of the U.S. government.

https://doi.org/10.1111/j.1754-9434.2009.01177.x Published online by Cambridge University Press

the position in question and consistent with business necessity.'' Thus, to say that a test such as Raven's Progressive Matrices is not job related when there is disparate impact is to say that the use of the test is unlawful. There is no problem, however, in saying that evidence for the job relatedness of this test is primarily determined by something other than test-job content match.

Conversely, it is possible to say that some selection consideration is job related (e.g., a ban against using methadone) without invoking validation and have it subsequently upheld in the courts. It is also possible to say with Guion (1978) that the method of test construction based on job-relevant content may provide justification for test use, without further empirical investigation or invocation of ''content validity.''

Murphy (p. 458) also asserted that a test that measured only one type of knowledge would not typically be seen as content valid, regardless of the relevance of that knowledge to the job. This seems to argue another kind of split between validity and job relatedness, but it is not clear who would maintain this. Presumably it is not the federal EEO enforcement agencies.[1] If the objective were to predict overall job performance, such a narrow focus might not make sense and the test would not be validly used for this purpose. But that would be the case regardless of validation strategy.

Content Considerations and Test Usage: A Tale of Two Test Batteries, Reconsidered

Murphy proposed a mail-room mix-up of the tests for entry-level machine operators and data entry clerks and posited that both test batteries would yield scores that were significantly and ''perhaps'' substantially correlated with measures of job performance. Most likely the correlations would not be markedly different than those that would be obtained had the mix-up not occurred. One might re-envision this situation with the state professional licensing exams for attorneys, psychologists, and physicians being mixed up so that no one got the intended exam. Murphy allows that it is ''possible'' that measures of very specialized job knowledge (e.g., certification tests for advanced medical specialties) ''might'' show more validity for their intended jobs. In this mix-up, one would think that a decision to let the test results stand because there is likely some (unspecified) degree of correlation among the tests would not be prudent. There would seem to be some limits on ignoring test content.

The model that would make results indifferent to which test is administered presumably depends on underlying commonality among the tests. Where the tests have general cognitive ability as this commonality, of necessity, the tests correlate with *g* and with each other. (Murphy, Dzieweczynski, and Zhang [2009] note this commonality; however, they also note that the effects of positive manifold can occur whether or not the tests measure the same construct.[2]) But let the machine operator test cover specialized job knowledge of operating a particular kind of machine where applicants can reasonably be expected to have acquired this knowledge, while the data entry test is a clerical ability battery for people with no specialized knowledge. The clerical battery, one suspects, would have value for predicting who (after training) could do either job. But if the objective is to select machine operators ready to operate those machines now, then the more reasonable strategy would be to test for knowledge proximate

[1] The policy of these agencies is still embodied in the *Uniform Guidelines on Employee Selection Procedures* (1978). In addition, there are questions and answers (Q&As) that were subsequently published and which are available at www.eeoc.gov/policy/docs/qanda_clarify_procedures.html. Q&A 93 is relevant to the testing of only one attribute. Q&As, together with the *Guidelines*, are also available at www.uniformguidelines.com.

[2] Murphy et al. (2009) state that the arguments on the effects of positive manifold apply to test batteries; content matching may be more important for individual tests. Also, while positive manifold may be the norm for many classes of selection tests, it is not likely to be present in all cases.

https://doi.org/10.1111/j.1754-9434.2009.01177.x Published online by Cambridge University Press

to the job, rather than cognitive ability to acquire knowledge. The test for machine operators likely would have limited use for selecting the data entry clerks.

Testing for data entry clerks presents an interesting combination of intercorrelation and content as well as introducing the legal dimension. In one large operation (state income tax data entry), a clerical battery was shown to correlate about .30 with job performance (net keystrokes per hour) and have adverse impact by race close to statistical significance. A content-based keying performance test (not a typing test!) came in around .60 and with no adverse impact. Both tests correlate with the criterion and with each other. Selecting which test to use is not a matter of indifference. Test content matters. These results were for key-what-you-see jobs. Jobs that require the clerk to encode information before data entry have other cognitive demands and a different testing strategy involving both cognitive and performance tests. Job content matters.

The Nature of Validity and Validation

Content considerations are an intrinsic aspect of construct validity. Unfortunately, over the years there seem to be at least as many variations on these considerations as there have been authors. Content may be limited to observable behaviors and outcomes, expanded to include operationally defined person attributes, expanded further to subconstructs and unobservable processes, or taken as the totality of the assessment event. The *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 11) include as test content ''the themes, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring.'' Content can refer to the domains of considerations deemed relevant or extraneous for defining predictor and criterion constructs and developing appropriate measures (Society for Industrial and Organizational Psychology, 2003, pp. 16–18). In this latter usage, content explicates what predictor–criterion relationship is being demonstrated by the validation effort. ''Valid for what'' is intrinsic to meaningful validation. That there could be various ''whats'' with some degree of intercorrelation does not negate this imperative and turn it into ''valid for whatever.''

Given the importance of content, we can well ask whether positive manifold (or other correlational analysis) alone establishes validity. Humphreys (1994), citing Meehl (1986), indicated that what positive manifold defines is a ''surface quasi-trait,'' something that depends on covariation among measures but has no obvious sources of similarity. Humphreys accepted dust bowl empiricism. But other people have a problem with content-free covariation.

Borsboom, Mellenbergh, and Van Heerden (2004), reviving the realist perspective on validity, maintained that the concept of validity expressed as ''the test measures what it purports to measure'' implies that there is an attribute that can be measured and this attribute causally affects measurement. Correlations among measurements can vary for reasons other than differences in the attribute purportedly being measured. Simply considering correlations alone could involve attributes that have no logical connection to the attribute of interest. Validity is not about such relationships, but the processes that link the attribute to variations in measurement.

What is interesting is not so much which formal theory is popular among psychologists but how formal theory sheds light on implicit validity theories of nonpsychologist stakeholders. One set of stakeholders is the people (attorneys, government investigators, labor economists, and human resources specialists) who are involved in the enforcement EEO law. There apparently has been no published study of their views, but anecdotal evidence suggests that there would be more concern with attribute causality than with nomological networks.

https://doi.org/10.1111/j.1754-9434.2009.01177.x Published online by Cambridge University Press

Murphy (p. 461) noted that, ''Arguments based on meta-analysis, principal components analysis, and other empirical principles are simply less compelling to most stakeholders than demonstrations of job relatedness.'' That statement is likely true, simply because job relatedness *is* the critical issue. Validation for EEO stakeholders involves concern for the test content mentioned above in the *Standards*, insofar as it impacts the meaning of test scores for legally protected classes. It also includes concern for why tests with equivalent validity differ in adverse impact (Outtz, 1998).

The Impact of Murphy's Thesis

''Content validity'' is likely to remain with us indefinitely. It is entrenched in our nomenclature and our practice. Call it ''validity'' or not, it is useful. However, it might be less used, and less misused, if test practitioners could avail themselves more readily of tests in which both statistical and content evidence exists. Two possible means of doing this are validation transportability and synthetic validity. Both approaches offer the extended use of correlation evidence beyond the original validation effort while avoiding the criticism of lack of job content in meta-analytic validity generalization.

Transportability depends on similarity between the situation in which validation occurred and a new situation for applying the validation results. For the federal EEO enforcement agencies, this requires having substantially the same major work behaviors. The comprehensive but daunting approach to establishing job similarity discussed by Gibson and Caplinger (2007) may be unnecessary if, as Murphy indicates, alignment of test and job content (here, test and new job content) means little for validity. Transportability's application is limited to situations in which fully validated tests exist for complete jobs. Synthetic validity, involving the assembly of tests shown to be valid for specific groups of job tasks into a battery for a whole job, would have more applicability. Here again Murphy's thesis would seem to negate the need to sweat the details on job content. Unfortunately for synthetic validity, it also highlights a problem. As long as we draw from limited domains of predictors and criteria, it may not much matter how test batteries are assembled. Scherbaum (2005) noted that batteries assembled through synthetic validity sometimes worked as well for unrelated occupations as for those occupations for which they were developed—exactly Murphy's point. Possible explanations included considerations that performance on job components was general rather than specific, or that cognitively slanted job analyses resulted in batteries primarily of cognitively oriented predictors. Murphy has indicated that the problem is likely a function of the intercorrelation of our measures. Scherbaum did not consider the lack of discriminant validity to be a fatal flaw for synthetic validity. Indeed, synthetic validity could facilitate the expansion of potential predictor and criterion domains. But it would seem that Murphy has provided vital clarification for what some issues are, and are not, if synthetic validity is to become a practice more used.

In Summary

Job relatedness is not just for the psychometrically unsophisticated or those who deal with legal issues in employment testing. The specification of test and criterion content is intrinsic to a meaningful concept of validity. Positive manifold does not supplant the place of content considerations in validation. That having been said, Murphy's contribution has been to bring focus on test-job content match concerns that are not really relevant to validity. The impact of this may not be so much in the traditional area of content-oriented test development, but for the newer and currently less used alternatives to classic criterion validation.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards*

https://doi.org/10.1111/j.1754-9434.2009.01177.x Published online by Cambridge University Press

for educational and psychological testing. Washington, DC: American Educational Research Association.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Civil Rights Act of 1991, 42 U.S.C. § 2000e-2 (1991).

Gibson, W. M., & Caplinger, J. A. (2007). Transportation of validation results. In S. M. McPhail (Ed.), *Alternative validation strategies* (pp. 29–81). San Francisco: Jossey-Bass.

Guion, R. M. (1978). ''Content validity'' in moderation. *Personnel Psychology*, *31*, 205–213.

Humphreys, L. G. (1994). An unrepentant author. *Psychological Inquiry*, *5*, 211–214.

Meehl, P. E. (1986). *Trait language and behaviorese*. In T. Thompson & M. Zeiler (Eds). *Analysis and integration of behavioral units* (pp. 315–334). Hillsdale, NJ: Lawrence Erlbaum.

Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 453–464.

Murphy, K. R., Dzieweczynski, J. L., & Zhang, Y. (2009). Positive manifold limits the relevance of content-matching strategies for validating selection test batteries. *Journal of Applied Psychology*, *94*, 1018–1031.

Outtz, J. L. (1998). Testing medium, test validity, and test performance. In M. D. Hakel (Ed.), *Beyond multiple choice* (pp. 41–57). Mahwah, NJ: Lawrence Erlbaum.

Scherbaum, C. A. (2005). Synthetic validity: Past, present, and future. *Personnel Psychology*, *58*, 481–515.

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

Uniform Guidelines on Employee Selection Procedures. (1978). 29 C.F.R. §1607.

https://doi.org/10.1111/j.1754-9434.2009.01177.x Published online by Cambridge University Press