*Industrial and Organizational Psychology*, **2** (2009), 469–474.


# Evidence of Content Matching Is Evidence of Validity

GEORGE C. THORNTON III
*Colorado State University*

Words mean a lot; words from a recognized authority such as Murphy mean even more. When such words are polemic, a strong reply seems warranted.

I fear that some of Murphy's (2009) words will obfuscate the discussion of validity of personnel selection practices and have dire consequences in employment discrimination litigation. Examples of words that raise my concern include:

- ''Content validation is useful for many things, but validity isn't one of them'' (p. 453).
- ''Comparisons between test content and job content . . . have little if any bearing on validity'' (p. 453).
- ''. . . assessments of content validity turn out to have little to do with the validity of these tests as predictors of job performance'' (p. 454).
- ''Content validation studies provide information about job relatedness, which is neither necessary nor sufficient for validity'' (p. 462).

I fear such words will cause confusion because of their juxtaposition with other words in the article:

- ''There is little doubt that content-oriented methods of validating tests are useful for establishing the job relatedness of selection tests'' (p. 453).
- ''Validity is indeed often substantial when the content of the tests matches the content of the job'' (p. 455).
- ''. . . these [content valid] tests usually perform quite well and are likely to be valid predictors of performance'' (p. 456).
- ''Reliable batteries of tests. . . whose content matches up with the content of the job will almost certainly turn out to be valid predictors of job performance'' (p. 458).
- ''A good match . . . probably enhances the acceptability, legal defensibility, the apparent fairness and reasonableness, and transparency of the tests'' (p. 462).

Combining these two sets of words would seem to lead one to conclude that a test can be job related based on content validity evidence and predictive of job performance but *not valid*. Based on what I understand to be the modern understanding of test validation and validity, this conclusion does not follow. I argue here that content validity evidence, along with other evidence, supports the inference that a battery of tests is valid for making personnel selection decisions. I am using the term ''valid'' and ''validation'' as articulated in the leading authority on test

Correspondence concerning this article should be addressed to George C. Thornton.
E-mail: George.Thornton@colostate.edu
Address: George C. Thornton III, Department of Psychology, Colorado State University, Fort Collins, CO 80523.



https://doi.org/10.1111/j.1754-9434.2009.01175.x Published online by Cambridge University Press

validation, namely, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & American Council on Measurement in Education, 1999). Surprisingly, Murphy does not cite this document. Although Murphy relies on the single set of evidence of positive manifold between tests and criteria to render content validity irrelevant to validity, I subscribe to the contemporary definition of validity that embraces content matching along with other types of evidence to support inferences about the validity of test scores.

## Historical Context

Landy (1986) admonished us to stop ''stamp collecting'' when validating tests. He likened claims about a test's validity to hypotheses that are tested by gathering varieties of evidence. There are NOT distinct and different types of validity labeled ''content validity,'' ''predictive validity,'' ''content validity,'' and ''construct validity.'' Rather, all evidence bears on the validity of a test. This means that we should not be trying to gather just one type of evidence to support one inference and another type of evidence to support another inference. All evidence that accumulates over time contributes to our understanding of a unitary concept of validity.

Murphy discounts evidence of content matching to support the validity of a test. By contrast, content matching is listed as one source of evidence in the *Standards*. The modern view of validation espoused by Landy and others became codified in the 1999 version of the *Standards*, which state ''Validation is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose'' (p. 457). Further, ''A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses.... Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system'' (p. 460).

The following types of evidence are enumerated in the *Standards*: test content, response processes, internal structure, relations with other variables including tests measuring similar constructs and criteria, and consequences of testing. As Murphy notes, matching test and job content has many positive consequences, for example, acceptability, perceived fairness, and transparency. In addition, content matching frequently *minimizes* one of the highly problematic unintended *negative* consequences in personnel selection, namely, adverse impact. Adverse impact against legally protected subgroups including racial/ethnic minorities, women, and older candidates is often substantially reduced with behaviorally based techniques that match job content. I argue that these consequences are directly relevant to the validation of a test battery.

Although Murphy lists a number of benefits of content validation, he appears to discount these other types of evidence in forming conclusions about validity: ''Selection test batteries made up of very different tests will usually show similar levels of validity, and there is little evidence that test batteries whose content matches the content of the job will in fact turn out to work any better than alternate batteries whose content is not matched to the job'' (p. 459). I contend that different test batteries have different levels of validity (i.e., ''work better'') when one considers the several intended positive consequences and avoidance of unintended negative consequences that derive from content validity tests. Murphy's thesis is that evidence of content matching is not evidence of validity. In contrast, the *Standards* embrace judgments of qualified subject matter experts regarding the probative value of content validity evidence in the validation of a test for personnel selection.

https://doi.org/10.1111/j.1754-9434.2009.01175.x Published online by Cambridge University Press

Professional judgments about the validity of a test involve a careful consideration of the variety of diverse evidence that has been accumulated about the test. We are long past the point when we are just ''stamp collecting'' (Landy, 1986) and looking for one type of evidence, namely, predictive correlations. A valid test is one with diverse supporting evidence.

## Professional Judgment on a Practical Example

Consider the sets of evidence for the following two test combinations. Which examination would professional judgment say is ''valid'' for making promotion decisions in a police or fire department?

- Promotional Examination A: high content similarity of test and job; one part has 100 multiple choice questions measuring knowledge of rules and regulations and another part with three simulation exercises calling for candidates to deal with situations encountered on job; candidate and managerial acceptance; no adverse impact; high face validity; lack of negative unintended consequences, that is, low probability of legal challenges and high probability of legal defensibility.
- Promotional Examination B: low content similarity; one part has abstract questions measuring general reasoning abilities (a test of *g*?) and another part has a self-report questionnaire calling for candidates to describe their leadership styles; low user acceptance; substantial adverse impact in the *g* test; correlations of scores with criteria; low face validity; negative consequences in the form of a high probability of legal challenges and high probability of legal indefensibility.

Some, including Murphy based on this article, might say Exam B is valid because of its correlation with criteria. Others would probably conclude that Promotional Examination B is *not* valid for making promotion decisions. I would endorse that conclusion. Despite the criterion-related validity evidence, other relevant evidence indicates this is not a valid test combination for this situation.

Rather, despite the *lack of* correlation evidence for Promotional Examination A, I would argue that other evidence about this test indicates it *is* valid for making promotional decisions. Murphy's analysis and conclusion asserts that evidence of representativeness of test content to the job domain is relevant to acceptability of the test, transparency, and legal defensibility, but is NOT evidence of validity. I disagree: Such evidence and its consequences support the inference that the test is valid for promotional decisions.

## Focus on Positive Manifold

Murphy focuses on evidence of positive manifold, the condition in which tests and criteria all have similar positive intercorrelations. This evidence is persuasive in showing that tests with low content matching can show test–criterion correlations (Murphy, Dzieweczynski, & Yang, 2009). But, the interpretation of these findings must be scrutinized. Just because Test B (which does *not* show content similarity to the job) can predict a criterion does *not* mean Test A (which *does* show content similarity to the job) is not valid.

Murphy argues that the condition of positive manifold imposes limits on the unique contribution of Test A. He gives inadequate consideration to the condition in which there is minimal correlation between two valid predictors. For example, promotional examinations for supervisory positions in many public jurisdictions include multiple choice tests of knowledge of rules and regulations and observations and judgments about overt behavior in simulations of key job situations (i.e., the assessment center method). In my experience with numerous promotional examinations in police and fire departments, the correlation of written tests of knowledge and assessment center

https://doi.org/10.1111/j.1754-9434.2009.01175.x Published online by Cambridge University Press

ratings correlate about .25. A variety of evidence suggests both these two measures are valid. Evidence of content matching comes from many sources, for example, job analysis, subject matter experts, thorough training and certification of assessors. Statistical studies of the relationship of test scores and criterion measures (so called predictive or concurrent validity) are not feasible because of the need for test security, small sample sizes, and the lack of adequate criterion measures. That does not mean there is no evidence of validity. Candidates who possess more knowledge of rules and regulations and better skills at decision making and management will probably be better lieutenants or captains. These tests are not only job related but also valid.

## Methods and Constructs

Murphy's analysis does not make adequate distinctions among different types of constructs and measures. His analysis focuses on paper-and-pencil measures of general cognitive abilities and self-report measures of personality. His mentions do not adequately consider work sample and simulation tests. In the throes of earlier debate about the value of content validity evidence, Tenopyr (1977) made the cogent point that one must consider the inferential leap that is being made when judging whether performance on a test would predict performance on a job, and thus whether content validity evidence is probative. In some cases, the inferential leap is great, for example, inferring that scores on abstract figural reasoning tasks such as Ravens Matrices relate to job performance. Thus, we must have evidence of statistical relationships. On the other hand, with other test formats measuring other constructs, the inferential leap is small: for example, inferring that scores on a work sample test of speed and accuracy of data entry is predictive of speed and accuracy of data entry on the job. Thus, evidence of content matching is probative; that is, we can conclude the test is valid for this purpose.

It is not a large inferential leap to assert that candidates who possess more knowledge of rules and regulations shown in a multiple choice test and better behavioral skills at decision making and management shown in simulations of relevant job tasks will be better lieutenants or captains. These tests are not only job related but also valid. That is, they can be validated by collecting evidence that support the inference that higher scores are related to job performance.

## Content Validity Evidence

The modern understanding of validation with the content validation process goes far beyond a superficial claim that the test looks like the job. It involves diverse bodies of evidence including thorough analysis of job tasks and requisite knowledge, skills, and abilities; careful test development (including items, other test stimulus materials, and instructions that match job requirements); scrutiny of test content to eliminate unfair content; demonstration of reliability including certification that assessors of behavioral observations show inter-rater agreement; confirmation by independent subject matter experts that test content and responses match the job; and so forth (Thornton & Mueller-Hanson, 2004). Thus, a variety of evidence is brought to bear.

The general dismissal of evidence of content matching as irrelevant to validation is too extreme. A more analytical dialogue, one in line with modern thinking about validity, would proceed like this: In what circumstances is evidence of content matching and other evidence, in the absence of a local study of test–criterion relationship, adequate to support the inference that test scores predict job performance and are valid for use in personnel selection? In like manner, in what circumstances is evidence of a test–criterion relationship, in the absence of evidence of content matching, adequate to support the inference? I hope the dialogue that ensues from this exchange of views will generate more analytical answers to these questions.

https://doi.org/10.1111/j.1754-9434.2009.01175.x Published online by Cambridge University Press

## Practical Considerations

Murphy's definition of validity would be highly injurious to the ability of organizations to conduct test validation and demonstrate the validity of selection and promotion systems. Consider this highly common situation: The civil service department of a medium-sized city develops biannual promotional examinations. This process is repeated every 18–24 months for the promotion of lieutenant, captain, and battalion chief in fire, and for promotion to sergeant, lieutenant, and captain in police. The city charter requires that promotions be based on merit and fitness. Past law suits, settlement agreements, and agreements with employee associations require that the promotional exam be reliable and valid.

The civil service staff, in conjunction with an industrial psychologist, conducts a thorough job analysis involving study of resource material, interviews job incumbents and supervisors, and administers questionnaires. They identify a number of attributes required for job performance including sets of knowledge and abilities. They design a multiple choice test of rules and regulations covering critical source documents and a set of behavioral exercises simulating critical job situations including an oral presentation, in-basket, and tactical exercise. These simulations afford the opportunity to observe behavior relevant to the abilities of oral communication, problem analysis, and leadership. Several secure Angoff-like panels establish the relevance and difficulty of test items. A secure panel of battalion and deputy chiefs establish that the content of simulations match job challenges.

Candidates are administered the written test on one occasion and the assessment center exercises on another date. Scores on the written test demonstrate acceptable internal consistency. Assessors at midlevels from police departments in comparable cities are trained and certified to reliably observe and rate behavior relevant to the performance dimensions in the assessment center. Interassessor agreement of ratings in the operational program is established. Scores on the written test and behavioral assessment correlate +.25 (the level of correlation between these two components has ranged between .25 and .35 in other comparable promotional programs in this city). The two methods are weighted and combined to provide a final promotional score.

Owing to security concerns and the practical pressures to develop promotional exams each year, it is not feasible to study the statistical relationship of test and assessment center scores before administration. It is not feasible to administer the measures to a candidate group and wait to gather subsequent criterion data, nor is it feasible to administer the measures to a concurrent group and gather current criterion data.

It is certainly not reasonable for the city to investigate if alternative tests which do not match the job are also valid. Is it plausible (likely) that a combination of a general mental ability test (*g*), a written test of judgment (e.g., Watson-Glaser), and a questionnaire measuring leadership style (i.e., tests that do *not* match the job) would show predictive validity? The principle of positive manifold would say ''Yes.'' Does that mean the city cannot assert that the promotional exam has validity? We, and the leading authority on test validation, that is, the *Standards*, would say the tests are valid. The words quoted at the beginning of this comment suggest Murphy would say the promotional exam is not valid, despite the fact that the exam is job related and would probably predict performance.

## Obfuscation of Legal Dialogue

Murphy says that evidence that the content of a test matches content of the job has ''little if any bearing on the predictive validity of selection tests'' (p. 453). Elsewhere he says that content valid tests are likely to be job related and will almost certainly be valid predictors of performance. These statements add considerable confusion, nay

https://doi.org/10.1111/j.1754-9434.2009.01175.x Published online by Cambridge University Press

obfuscation, to the already difficult challenge of providing expert witness testimony in employment discrimination litigation. In the language of courts, expert testimony must be relevant and reliable (Thornton & Wingate, 2005). In this context, relevant means the testimony is directly pertinent to the specific situation being litigated. Reliable means the testimony meets standards set forth in rules of court evidence. In this context, reliable evidence is equivalent to the psychometric properties of reliability and validity. To say that content validity evidence doesn't support the validity of a test is not only confusing but also damaging.

Is it plausible that a test is ''job related'' but not ''valid?'' Maybe in some new definition of validity, but even that is hard to defend and explain. Murphy's definition of validity would preclude the presentation of evidence of content matching to show the validity of a selection test. Imagine the opposing expert or attorney saying ''Well, Dr. Guion (former president of SIOP and editor of *JAP*!) says there's no such thing as content validity. And, Dr. Murphy (former president of SIOP and editor of *JAP*!) says that content validity evidence is not evidence of validity. Why do you say it *is*?''

The expert witness would be faced with explaining why these experts contend that content-matching evidence (i.e., content validity evidence) is not germane to the validity of the test and then discounting that contention. No small challenge in the press to be concise in front of judge or jury in a court room.

## Summary

I suggest that addressing the challenges faced by Murphy's assertions and answering the attorney's questions, one can say: ''The leading authoritative documents in the field have concluded that Guion (1997) was wrong 30 years ago: Content validity evidence does exist, and it is useful in test validation. And, Murphy is wrong now today: Evidence of the match between test and job content IS evidence of validity for many tests of many characteristics important for job performance. We must note that Murphy also said that content valid tests perform quite well and will probably turn out to be valid predictors of job performance.''

## References

American Educational Research Association, American Psychological Association, & American Council on Measurement in Education. (1999). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

Guion, R. M. (1977). Content validity–The source of my discontent. *Applied Psychological Measurement*, *1*, 1–10.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*, 1183–1192.

Murphy, K. R., Dzieweczynski, J. L., & Yang, Z. (2009). Positive manifold limits the relevance of content-matching strategies for validation selection batteries. *Journal of Applied Psychology*, *94*, 1018–1081.

Murphy, K. R., Dzieweczynski, J. L., & Yang, Z. (in press). Positive manifold limits the relevance of content-matching strategies for validation selection test batteries. *Journal of Applied Psychology*.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology*, *30*, 47–54.

Thornton, G. C., III, & Mueller-Hanson, R. A. (2004). *Developing organizational simulations: A guide for practitioners and students*. Mahwah, NJ: Lawrence Erlbaum.

Thornton, G. C., III, & Wingate, P. H. (2005). Industrial and organizational psychologists as expert witnesses: Affecting employment discrimination litigation post Daubert. In F. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 167–198). San Francisco: Jossey-Bass.

https://doi.org/10.1111/j.1754-9434.2009.01175.x Published online by Cambridge University Press