Industrial and Organizational Psychology, **2** (2009), 486–492.

Coherent Conceptualization Is Useful for Many Things, and Understanding Validity Is One of Them

JOHN F. BINNING
The DeGarmo Group, Inc., and Illinois State University

JAMES M. LEBRETON
Purdue University

Murphy (2009) contends that "... *assessments of ... the content of tests and the content of jobs turn out to have very little to do with the validity of tests as predictors of future job performance*" (p. 455). He goes on to conclude that "*(c)ontent validation does not tell you much about validity ...*" (p. 460). In drawing these conclusions, he confounds the conceptual role of content validation with attempts to subject that role to empirical verification. In addition, the empirical base for his conclusions is notably insular. As a result, his criticism of content validation is unwarranted. By adopting a myopic view of content validation, focusing on one specific issue of how test batteries are assembled, and confounding predictor constructs with predictor methods, he seemingly creates a straw man argument that could cause considerable confusion for those interested in a clear understanding of validity. Our goal in this commentary is to clarify and reinforce the integral role that content validation should play in sound theory development and credible applications of psychological science. In addition, we hope to delimit the implications of Murphy's claims about validity so that they are not misinterpreted or inappropriately generalized.

Evidence of validity based on test content (i.e., content validation) is a much broader, more conceptually and empirically intricate process than Murphy characterizes. Psychological science is built on (a) hypothesizing the existence of and identifying domains of covarying behaviors, (b) specifying the boundaries and internal structure of these behavioral domains, and (c) theorizing about their causes, effects, and inter-relationships. These scientific activities are intimately intertwined with content considerations. Applying this science to employment decision making requires extrapolating from theory about general behavioral regularities to specific workplace circumstances. In a sense then, validity refers to the extent to which domain specification and extrapolation processes are carried out in a scientifically credible manner, and content validation is the core of those processes. Although the *Principles* (2003) succinctly define validity as "(t)he degree to which accumulated evidence and

Correspondence concerning this article should be addressed to John F. Binning.
E-mail: jbinning@ilstu.edu
Address: The DeGarmo Group, Inc., and Illinois State University, 2116 S. Morris Ave., Bloomington, IL 61704
John F. Binning, The DeGarmo Group, Inc., and Illinois State University; James M. LeBreton, Department of Psychological Sciences, Purdue University

https://doi.org/10.1111/j.1754-9434.2009.01178.x Published online by Cambridge University Press

theory support specific interpretations of scores from a selection procedure entailed by the proposed uses of that selection procedure'' (p. 72), we believe a systems perspective best typifies how content validation relates to validity.

A Systems View of (Content) Validity

In the employment decision context, validity is a property of a (complex) adaptive information system, often comprising a blend of actors, contexts, and activities that operate and fluctuate over time. To say that a selection system is valid is to say that there is convincing evidence that a system of information gathering (e.g., predictor sampling) and data integration (e.g., decision rules and interpretive judgments) works the way it was designed to work, and that it systematically produces predictions of work behavior that conform to an intended purpose. There is no specific amount or type of evidence that guarantees a particular system will be viewed as valid by any particular individual in a given situation, but a preponderance of evidence suggesting such, generally can be recognized by appropriately knowledgeable people.

The core content of an HR selection system is information gathering and integration processes, specifically, (a) job or work analyses to identify criterion construct domains (CDs) (also known as performance domains) of situationally valued work behaviors or outcomes to be predicted, (b) a behavior sampling process used to gather information upon which to base predictions, and (c) a working theory about how predictor information should be integrated to make the intended predictions. A system can be relatively simple and manageable (e.g., a single mechanically scored structured test and top-down decisions) or complex and logistically burdensome (e.g., a multistage, diverse battery of structured and unstructured information gathering and scoring methods, and nonlinear, configural decision models). Regardless of the complexity, a selection system is valid to the extent that evidence supports the conclusion that the design and operation of the constituent processes conforms to relevant standards (e.g., produces accurate decisions, is sufficiently job related, and consistent with business necessity). From this systems perspective, it is easy to see that validity is not a dichotomous state, despite the fact that professional and judicial review often yield what appears to be such a determination. The validity of a selection system is an ever-evolving *system state* in that many system variables are interacting, and at any point in time, these can be thought to be arranged in some particular configuration (e.g., likely to produce valid selection decisions). Given the dynamic complexity of the system, one can only infer validity from available evidence. It can never be determined with complete certainty.

Consider the parallels between evaluating a selection system's validity, and deciding on your physician's professional competence, determining the condition of a prospective automobile purchase, or evaluating a politician's reputation when deciding how to vote. Confidence in these ''systems'' covaries with the availability of confirming and disconfirming information. One's confidence increases incrementally as information consistent with expectations unfolds but can decrease precipitously, given dramatic new information (e.g., a discovered lapse in selection system administration, a sensational malpractice claim, a CARFAX flood damage report, or the public revelation of an extramarital love child). In any particular time frame or circumstance, the nexus of all available evidence leads any interested party to evaluate the system state's acceptability for a particular purpose, with varying degrees of confidence. Therefore, validity is ALWAYS to some degree a matter of faith—faith that predicted outcomes in the future will mirror those for which evidence was specifically compiled in the past. A very important source of this faith is the quality with which the content of the predictor sample is specified, designed, and administered.

Validity is a quality of the system relative to the decisions it produces. The

https://doi.org/10.1111/j.1754-9434.2009.01178.x Published online by Cambridge University Press

component processes produce information that is used in a particular way to produce decisions. Thus, validity is dependent not only on how information is gathered but also how that information is used. Put another way, psychological constructs (e.g., cognitive ability) and behavior sampling methods (e.g., biodata forms, interviews, work samples, or situational judgment tests) do not predict anything or do criterion constructs. These are merely components of an information system that if guided by an appropriate theory of how to use available information can predict important outcomes. Of course, the specific outcomes are relevant, as well, because the same information can be manipulated differently depending on what is being predicted. Recall the Connecticut police department that disqualified police candidates who scored too high on an intelligence test. Their theory of predictor–criterion relations suggested that intelligence was monotonically related to the likelihood of experiencing boredom and voluntarily terminating employment. Change the criterion (e.g., report writing proficiency; investigative skill, etc.), and the same predictor information would likely be used differently.

Viewed this way, validity is a summary evaluation of the prototypical operation of a selection system, and myriad forms of evidence are relevant for characterizing the quality of this operation. The history of how to categorize, conceptualize, and label these evidential bases is the focus of much scientific and professional discussion over the years (cf. American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1985, 1999; Binning & Barrett, 1989; Society for Industrial and Organizational Psychology, Inc., 1987, 2003). Regardless of the specific category labels, the core conceptual substrate of validation has very much to do with predictor content. We will now focus on how the content of a predictor process is essential to validity.

The Two Faces of Content Validity

There is a long-standing reliance on content considerations in the technical and professional guidelines for psychological testing, most notably the *Uniform Guidelines* (*Uniform Guidelines on Employee Selection Procedures*, 1978), the *Standards* (1985; 1999), and the *Principles* (1987; 2003). These latter two documents are periodically revised in attempts to codify prevailing views of validity, and while having undergone some dramatic conceptual and semantic changes in the past several decades, they still emphasize the importance of content considerations in the validation of employment decision systems. The evolved change most relevant to this commentary is the deletion of any references to ''construct validity'' or ''content validity'' and the melding of both into ''content-based validity evidence'' or ''evidence based on test content.'' This has implications for how to conceptualize content validation. The *Principles* (1999; and *Standards*, 1999) state that:

> *Test content includes the questions, tasks, format, and wording of questions, response formats, and guidelines regarding administration and scoring of the test. Evidence based on test content may include logical or empirical analyses that compare the adequacy of the match between test content and work content, worker requirements, or outcomes of the job* (p. 6).

Although the list of constituent content could be elaborated, it is clear that test (predictor) content is broadly conceived. It is also clear that content validation involves matching, via logical and empirical analyses, test content to different job-related construct systems (i.e., work content, work requirements, and/or work outcomes). So what is involved in matching predictor content to work content versus work requirements versus work outcomes?

https://doi.org/10.1111/j.1754-9434.2009.01178.x Published online by Cambridge University Press

The melding of traditional construct and content considerations means that predictor content is conceptualized as sampling either from criterion CDs (i.e., work content or work outcomes) and/or psychological CDs (i.e., worker requirements). Criterion CDs are clusters of task situations, activities, and outcomes induced to covary by organizational imperatives, typically delineated via job and work analysis, and labeled in the language of the organization (e.g., customer service orientation). Psychological CDs are clusters of naturally covarying behaviors delineated by psychological research, representing scientific tools of a more general individual difference psychology, and labeled in the language of the science (e.g., agreeableness and extraversion). Thus, content validation involves either (a) directly matching predictor content to criterion CDs or (b) matching predictor content to psychological CDs which are in turn related to criterion CDs (i.e., delineating psychological traits believed to influence job behavior). Regardless of whether (a) or (b) is targeted, predictor content includes not only behaviors but also the parameters within which those behaviors are assigned meaning, such as the specific stimuli that evoke behavior, the instructional and administrative environment in which behavior is evoked, the format in which behavioral information is recorded, and how behavior is interpreted. However, (a) and (b) involve different conceptualizations of content involving different domain specification processes and different inferential linkages. To facilitate a discussion of these distinctions, an adaptation of the diagram appearing on p. 153 of the *Standards* (1999; and p. 104 of Guion, 1998) is presented in Figure 1.

Predictors sample criterion CDs (in Figure 1) or psychological construct domains (ΨDs in Figure 1). Validation involves generating and/or compiling evidence to support Inference 5 in Figure 1—the inference that from predictor information, future job behavior, and outcomes can be reliably

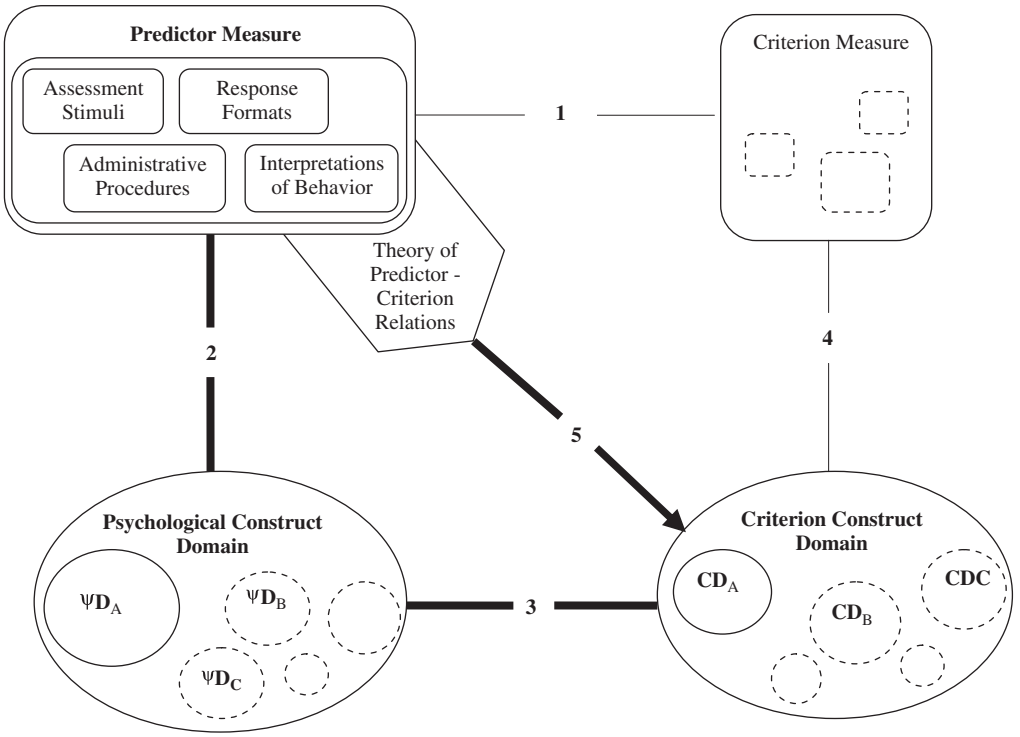

Figure 1. Inferential linkages involved in content validation.

https://doi.org/10.1111/j.1754-9434.2009.01178.x Published online by Cambridge University Press

predicted. Figure 1 depicts the component processes of a selection system, and therefore, the various foci of validity evidence. *A crucial point for the current discussion is that content validation of a particular predictor process will differ depending on whether it is designed to tap criterion construct domains versus psychological construct domains.*

Psychological Constructs and Content Validation

Content validation of a predictor process designed to tap psychological constructs involves garnering evidence to support BOTH Inferences 2 and 3. Inference 2 is supported by evidence that a given predictor adequately samples from a specific psychological CD. This adequacy is highly dependent on the quality of the theory surrounding the target construct, especially the explicitness of its behavioral domain. The adequacy of this content sampling is ideally based on deep nomological understanding not merely on nominal or surface considerations.

Imagine that you apply for a job as a financial collections representative and are connected to an EKG monitor to measure various spacing parameters between your heart beats. Now consider the content validation of such a predictor. It might go something like this. There is evidence that higher levels of self-esteem inoculate employees from the emotional duress caused by contentious client interactions (i.e., evidence for Inference 3 relating ΨD_A to CD_A). There is also evidence that vagal tone—the tonicity of the vagus nerve in the parasympathetic nervous system that innervates the major internal organs—reliably correlates with self-esteem (e.g., evidence relating ΨD_A to ΨD_B). And finally, there is evidence that vagal tone can be reliably measured by EKG monitoring of certain heart beat (e.g., respiratory sinus arrhythmia) parameters (i.e., evidence for Inference 2). If an employer chooses to use this measure of vagal tone as a predictor, an important part of the validity argument would focus on the specific content of the vagal tone CD. This argument would be strengthened by presenting theory and data on how parasympathetic nervous system neurons in the heart determine vagal tone in the service of threat response and how this is thought to underlie state self-esteem. Additional content considerations would include the administrative environment in which EKG measurements are taken, the format in which EKG information is recorded, and how EKG readings are interpreted. Our point is that the content validation of a predictor designed to tap psychological constructs requires deep theory about CDs, and content considerations lie at the (pardon us) heart of these various content validity arguments, traditionally labeled ''construct validity.'' In summary, content validation of predictors designed to sample psychological CDs is conceptually deep, serious business that must be viewed that way, or one runs the risk of drawing simplistic conclusions that hamper both research and practice.

This is not a problem unique to HR selection science. Psychological science, in general, has been perennially challenged to adequately explicate the constructs that comprise its theories. From Agreeableness (or is that Interpersonal Effortful Self Control?) to Id (has it ever been measured?) to Intelligence (or are there several of these?) to Neuroticism (or is that Negative Affectivity?) to Passive Aggressive Personality Disorder (will it be in the DSM again?), the fundamental challenge is specifying the content of psychological CDs. Even a venerated charter member of the industrial–organizational (I–O) psychology construct lexicon (i.e., job satisfaction) was recently shown to suffer from serious domain specification problems (Weiss, 2002), which has in turn limited prediction and theory development. The central issue here is that psychological CD specification, done well, is arduous, even tedious work, and generally not as much fun as developing measures of one's favorite new constructs or theorizing deeply about how constructs relate to each other. Substantiating Inferences 2 and 3 requires

https://doi.org/10.1111/j.1754-9434.2009.01178.x Published online by Cambridge University Press

much more than comparing the name of a test to the name of a construct, and yet this name-matching process is often considered content validation. Unfortunately, we believe it is this form of content validity that Murphy invokes in his descriptions of research on cognitive ability.

Criterion Constructs and Content Validation

In contrast to the content validation described above, content validation of a predictor process designed to tap criterion constructs involves garnering evidence to support Inference 5 in a more inferentially direct way. Rather than garnering evidence of Inferences 2 and 3 to substantiate Inference 5, these inferences are in essence collapsed together, and the evidence based on content takes the form of substantiating the extent to which the predictor adequately samples criterion CDs. The ''theory'' about how to combine information to make predictions is driven by concerns for structural isomorphism between the predictor sample and actual job situations. Many consider this predictor–criterion ''resemblance'' to range from virtual isomorphism (e.g., long probationary periods) to work samples or simulation exercises. Regardless, the underlying logic is that adherence to gestalt principles will ensure predictive accuracy.

The same core content issues are relevant as before, but the content specification process is not as dependent on general psychological theory. Rather, the theoretical focus is on criterion CDs and how they are best reproduced in predictor samples. Concerns about content sampling shift from the adequacy of sampling psychological CDs to the adequacy of sampling criterion CDs. Remember that criterion CDs are clusters of *workplace behavior in situ*. Drawing on the financial collections representative example, the focus here would be on sampling task domains judged integral to collections representative performance (e.g., sampling important features of a recurring job situation requiring the use of verbal jujitsu to disarm an obstinate client). It is important to realize that self-esteem may well be operating in this situation, but the conceptualization of behavioral regularity is different from this perspective. As before, content validation in this context is focused on evidence that the predictor includes all of the processes associated with assessment stimulus presentation, administration, data collection methods, and assignment of meaning to the behavioral data. However, in this instance, these components will look *different* than those discussed for psychological constructs because *the behavioral domain being sampled is different*.

Especially relevant to this commentary is the fact that Murphy bases his criticism of content validation on three streams of research, which he claims to examine the relevance of content matching to validity. It is helpful to realize that research in his ''Weak Evidence . . .'' section exemplifies content validation focused on sampling *criterion constructs*. His criticism is weakened by mixed results and methodological confounds in the one study offering the ''most direct test of the hypothesis'' (e.g., CVRs were not computed on actual interview questions, restriction of range in validity coefficients). Research in his ''Large-Scale Tests . . .'' section focuses on content validation of *psychological constructs*. The fact that he focuses solely on cognitive ability, and matching only at the nominal level, serves to weaken his criticism. His ''The Structure of Selection Tests. . .'' section focuses on how the intercorrelations of tests in a battery can influence their statistical contribution to prediction. Again, the content matching is nominal in nature and even goes so far as to equate predictor components based on the name of a METHOD rather than CDs (i.e., ''measures of job-related knowledge, skills, and abilities to structured interviews, work samples, situational judgment tests, and even biodata inventories,'' p. 458). This clearly has little to do with content validation as we described above, and thus, further weakens his criticism.

https://doi.org/10.1111/j.1754-9434.2009.01178.x Published online by Cambridge University Press

A Closing Comment

Murphy closes by supporting his criticism of content validation with a quote from one of the clearest thinking, most influential validity experts, anywhere. He quotes Robert Guion from a 1978 paper in which Guion said there is ''... no such thing as content validity'' (Guion, 1978). We believe this is another example of Murphy relying on rather shallow conceptualization to support his claims. In the 1978 article, Guion was clearly admonishing those who reified ''content validity'' as part of the dissociative trinitarian view of the day. Guion was not suggesting clearly that domain content considerations were irrelevant to validity. This is directly corroborated by several claims in his 1978 article such as, ''(t)he comments made here do *not* argue that all employment testing requires empirical validation research,'' or ''(g)iven all this care and competence in the construction of the test, its use would certainly be justified without further research'' (p. 210). We believe that he was reacting to the widespread confusion and rampant conceptual imprecision with which terms like content validity were commonly being used 30 years ago. Our interpretation is corroborated by Dr. Guion's recently indicating that, ''The quotation was an intentional put-down of people who were quite willing to call a test valid (specifically content valid) simply because its content slightly resembled a larger domain of interest, without considering any other evaluative information or evidence'' (personal communication, June 17, 2009). We wholeheartedly concur that surface considerations of content do not constitute (content) validation. Since Murphy's conclusion rests largely on such a conceptualization of content validation, we must disagree with his conclusion. We do, however, agree with Murphy that, as usual, Guion was right!

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (Joint Committee). (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478–494.

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

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel selection*. Mahwah, NJ: Erlbaum.

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

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Society for Industrial and Organizational Psychology, Inc. (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.

Weiss, H. M. (2002). Deconstructing job satisfaction: Separating evaluations, beliefs, and affective experiences. *Human Resource Management Review*, *12*, 173–194.

https://doi.org/10.1111/j.1754-9434.2009.01178.x Published online by Cambridge University Press