Industrial and Organizational Psychology, **2** (2009), 465–468.

COMMENTARIES

Was This Trip Necessary?

ROBERT M. GUION
Bowling Green State University

Gasoline rationing during World War II required a sticker in the lower left corners of car windshields. Outside it identified the rationing level. Inside, it asked, ''Was this trip necessary?'' Murphy (2009) suggests a paraphrase of that old question: ''*Why* was this article necessary?'' I have two answers. First, unlike gas rationing, the 1940s notion of validity (expressed as a validity coefficient) has not yet gone away. Second, even the 1950s notion of validity is still not understood or accepted. To whatever extent I may be right, these are regrettable answers, and Murphy is not to blame.

Up through the 1940s, industrial psychologists understood validity as a property of a test and asked about test validity. We were not consistent. If validity inherently resided in a test, validity would have been as consistent across situations as validity generalization now says it is. We contradicted ourselves and insisted on situational validity, defined by correlations between tests and criteria. It didn't much matter whether the two measures were separated in time or, if they were, which measure was obtained first. Whatever the order or time gap, we would say we were ''predicting'' the criterion.

Early in Murphy's paper, we encounter this: ''Content-oriented validation strategies are flawed in the sense that assessments of content validity turn out to have little to do with the validity of these tests as predictors of job performance.'' This is the core of my question: Why should he have to say so at this late date? He's right, but a chain of psychometric giants (including a generation of people like Lee Cronbach, Ted Cureton, Lloyd Humphreys, Bob Linn, Sam Messick, etc.) repeatedly said so and were not heard—possibly not even read—by many personnel testers.

I find it hard to see how the successive versions of the *Standards,* from 1954 onward could be interpreted as suggesting that something called ''content validity'' could be expected to serve the purposes of something else called ''criterion-related validity.'' Maybe a version-by-version long view of *Standards* development would help.

It started in the 1950s with publication of the APA *Technical Recommendations for Psychological Tests and Diagnostic Techniques.* Following tradition, the article continued to refer sometimes to the ''validity of tests,'' but it also added the unnoticed but more important notion of the validity of hypotheses. It related validity to the ''four aims of testing,'' ending up with four ''types'' of validity: (a) content validity to estimate current performance in a ''content universe'' sampled by the test, (b) predictive validity to predict future performance on something else, (c) concurrent validity to estimate the person's status on something else at the time of testing, and (d) construct validity to infer the examinee's level of a presumed trait or quality (American Psychological Association, American Educational

Correspondence concerning this article should be addressed to Robert M. Guion.
E-mail: rmguion@wcnet.org
Address: 632 Haskins Road, Bowling Green, OH 43402

Research Association, & National Council on Measurement Used in Education, 1954, pp. 13–14). The ''four *aspects* of validity'' (APA et al., 1954, p. 13) were tied to these aims. Not until the 1966 *Standards* were the two ''types'' expressible with validity coefficients combined as criterion-related validity (American Psychological Association, American Educational Research Association, & National Council on Measurement Used in Education, 1966). (I asked Cronbach why the two had been combined; he answered that he couldn't see any good reason for concurrent validation.) The 1966 version, and later ones as well, avoided the term ''types,'' but both the 1954 and 1966 versions caused personnel testers to scramble to figure out what content and construct validity meant in personnel testing.

Suppose a universe of arithmetic problems varying widely in complexity; also imagine an arithmetic test that samples that universe very well. By a 1954 definition, the content validity of the test would be assured if ''very well'' meant a representative sample of components of the entire universe. The ''content-valid test'' might not be relevant to a lot of jobs, so scores on it would not predict performance of taxi drivers, plowers of corn fields, or astronauts—unless descriptions of those jobs would reveal complex but basic arithmetic skills of which I am unaware. They might; but it would require a data-based validity argument I, for one, am not prepared to offer.

I'll stay away from contrived, simplistic examples and go on with the 1966 version. It asserted that ''The three concepts of validity [note the change from *types* to *concepts*] are pertinent to all kinds of tests'' (APA et al., 1966, p. 14). That is, some element of test content has been relevant at least since 1966 to the validity of any kind of test (score) intended as a predictor of future performance. Murphy has it right; content validity (if there is such a thing) is relevant to criterion-related validity but was never thought to be a substitute for it by writers of the *Standards*.

The literature in psychological and educational measurement was extensive after 1966, much of it concentrating on defining validity. Drawing on that literature, the 1974 *Standards* said, ''Questions of validity are questions of what may properly be inferred from a test score'' (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1974, p. 25). The shift from the validity of tests to the validity of inferences was complete. It was not, however, universal; Anastasi (1986) argued that the validity of those inferences depended on validity built into the test, a point to remember. A competently developed test is more likely to permit valid inferences from its scores than one sloppily thrown together.

By 1985, the concept of validity was clearly separated from tests to the ''appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores'' (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985, p. 9), and reference to types, aspects, or concepts validity firmly gave way to the requirement of many sorts of evidence, of which the three labels were merely convenient bins into which a given piece of evidence might be stashed. In 1999, validity was again explicitly a property of inferences or interpretations from test scores (not a property of tests), and the holy psychometric trinity was fully abandoned in favor of more and different categories in which to pigeon-hole validity evidence.

The currently final word (as of 1999) is that ''Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests'' (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). Evidence of validity so defined may be based on content, response processes, internal structure, relations to other variables, and, although limited, on consequences of the testing process. Moreover, each of these

https://doi.org/10.1111/j.1754-9434.2009.01174.x Published online by Cambridge University Press

bases (which could be subdivided further) needs to be integrated into a solid *validity argument* explicitly relevant to the intended purpose of the testing. The notion of a validity argument did not spring on us as a big surprise. Cronbach (1988) insisted that validation accumulates a pile of research results—and good, reasoned judgment—to justify (or to reject) an intended inference, and it then attempts to arrange the evidence and judgments into a logical argument that might justify a particular use or interpretation of scores. If the attempt fails, the assumption of validity is not tenable. As Murphy makes clear, a content match between predictor and job makes a pleasing validity coefficient more likely, but I wish he had gone further to say that the match is part, but only part, of an overall argument supporting the predictor's use. Indeed, he seems to condone and continue referring to ''validation methods'' (referring, I think, to the traditional trinity of validity ''types'') as if they were entities on their own rather than pieces of a more comprehensive argument.

The *Technical Recommendations* started the whole thing by introducing those subsequently reified (deified?) words. The 1954, 1966, and the 1974 versions all insisted that these adjectives in front of the word validity merely identified *aspects* of validity. Unfortunately, they devoted so much space to the aspects of validity that they didn't get around to getting a good definition of what they meant by validity; validity per se became the great mystery in the psychometric sky. It still lacks a good coherent definition (at least for me), although the 1985 and 1999 versions made good stabs at getting rid of the troublesome word *aspects*.

When a word gets troublesome, draw a diagram. It will pull attention away from the word. So think of (and draw) a circle marked off in three segments, with the wedges separated by something like dotted lines to suggest the permeability and tentativeness of the segments. We can also think of the marks as moving (easy enough on a computer monitor but not on a printed page) to suggest *changes* in relative wedge sizes. The circle can represent the indefinable notion of overall validity (''of the test''?); the wedges can represent the relative importance of the aspects of validity for a specified circumstance. Under some circumstances, the content wedge will be the most salient; under others, perhaps the construct wedge is more salient. Under some circumstances, the evidence that a well-defined content domain has been well sampled is major support for claims of validity. These circumstances are constrained by the degree to which a content domain can be well defined; for some of Murphy's examples, it would be very hard to identify the edges of a universe or even a defined domain. Under other circumstances, the support that counts is statistical. Change the term *circumstance* to *intended use*, and change *support* to *evidence for*, and you have used the terminology of the 1999 *Standards* (American Educational Research Association et al., 1999). Welcome to the 21st century.

But note that 21st century terminology developed over a period of 4½ decades of grappling with the idea of validity and how it might be supported. In that time, test users should have learned that there are many ways to understand validity, that the ways are less than interchangeable, and that they are related to test *scores*, not to the *tests* from which the scores are derived. (But remember Anastasi!)

I'm sure that many will see this commentary as mere quibbling over semantics. I think the use of language is more important than that. Sometimes we use inaccurate or poor word choices as a kind of shorthand that lets us avoid using longer phrases that may be more precise. I have no problem with that. I don't even cringe very much when I hear talk about ''the validity of the test'' if I'm sure from other language used that the speaker knows that use of the term is merely shorthand. More often, however, I think the choice of words reflects the quality of thought, and if thinking about validity

https://doi.org/10.1111/j.1754-9434.2009.01174.x Published online by Cambridge University Press

as a property of test inhibits consideration of those things that might influence interpretations of scores, then the semantic distinctions are important to make and to keep.

Many psychometric scholars have hammered out the themes that there are different kinds of validity evidence (as opposed to different kinds of validity) and that an acceptable argument should be mustered to support an assertion of validity. More than one kind of evidence can and should be sought (Landy, 1986). No one piece of evidence makes or breaks the validity argument. To support (or fail to support), an argument of validity depends on the evidence in the main (''the preponderance of the evidence'' in legal jargon). Establishing the validity argument for predictive purposes is easier with good criterion-related validation, but even a superb meta-analytic mean validity coefficient (corrected, of course) is strengthened if other evidence (such as matching content) can be added to the mix. All of this should, I think, be considered well established by now.

We have, or should have, gone far beyond the 1940s assumption that a validity coefficient is the only necessary kind of evidence to support an inference of validity as defined in 1999. The sooner employment testers get up to date in their understanding of the many implications and nuances of validity evidence, the sooner authors like Murphy (and me) can stop chiding people for persistently wrong or misguided ideas. Murphy's point is right: The idea that matching test content domain to the content of task performance (or the match of constructs) is enough evidence that the scores will predict performance is really pretty stupid. My main quarrel is that, in making the point, Murphy perpetuates the archaic language that leads to such an idea.

References

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

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

American Psychological Association, American Educational Research Association, & National Council on Measurement Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51, 201–238.

American Psychological Association, American Educational Research Association, & National Council on Measurement Used in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–15.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

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

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.

https://doi.org/10.1111/j.1754-9434.2009.01174.x Published online by Cambridge University Press