Pesticide Applicator Certification Tests: Validation and the Limits of Score Meaning

Andrew Martin, Training Specialist, Purdue Pesticide Programs, Purdue University Cooperative Extension Service, West Lafayette, IN, martinag@purdue.edu

Fred Whitford, Coordinator, Purdue Pesticide Programs, Purdue University Cooperative Extension Service, West Lafayette, IN, fwhitford@purdue.edu

Abstract

A review of the basic elements of modern validity theory and an argument-based approach to validation clarifies the principle reason for a deliberative, job-oriented approach to pesticide applicator certification test development: appropriate score interpretation and use. When more meaning is imputed to scores than is warranted, stakeholders may be misled and program credibility can suffer. Special care should be taken to avoid making predictive claims for certification test scores. Caution is also advised when associating the concept of "competence" with score results.

Keywords: test, validation, scores, pesticide, applicator, certification

Introduction

Why is the concept of validity in testing so vital? The answer is that validity speaks to score meaning and is invoked to justify how tests get used. Curiously though, "For a concept that is the foundation of virtually all aspects of measurement work, it seems that the term validity continues to be one of the most misunderstood or widely misused of all (Frisbie, 2005, p. 21)." To the extent that misunderstanding or misuse of the term validity occurs in the pesticide applicator certification and training community, we run the very real risk of misleading ourselves and program stakeholders about what test scores really mean and what our programs actually do.

This paper examines modern validity theory and offers pesticide regulatory agencies a perspective on how to view their test construction activities. It addresses why we can only assign limited meaning to our certification test scores, and why it is inappropriate to make predictive claims based on score results or to attach ill-defined, value-laden labels such as competent and not competent to test takers based on their test scores.

Test Score Interpretation

You may have heard that validity is the extent to which a test measures what it purports to measure, or that there are different kinds of validity that are specific to various testing purposes. This should be reconsidered in light of a newer conception of validity (Angoff, 1988). Validity is currently recognized as a judgment of the extent to which evidence and theory support appropriate test score interpretation and use (Messick, 1989). Validity, properly understood, is not a test property; rather, it is associated with the interpretations that we assign to

test scores. This perspective permeates the current *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

A test score interpretation explains score meaning (Kane 1992). For pesticide applicator certification and licensing managers, a proposed score interpretation might sound something like this: “Test scores indicate achievement in important, job-related knowledge (e.g., knowledge of signal words) and skills (e.g., skill at arithmetic) necessary for entry-level practice.” What evidence, and how much evidence, would support this interpretation? Important preliminary evidence is derived from the test construction practice itself. If the test is simply a collection of questions written by a handful of specialists, as is often the case, it might reflect important job knowledge and skills, but that claim amounts to little more than an appeal to authority. A more robust, evidence-based claim rests on an organizational model for developing credentialing tests that entails:

1. Conducting a job analysis, for example by identifying worker activities and qualities that are necessary for effective job performance.
2. Developing a test plan based on the job analysis, by specifying testable knowledge and skills.
3. Assembling the test, by writing test items that reflect the test plan.
4. Determining test administration policies and practices, for example by taking steps to ensure fair treatment of all test takers.
5. Analyzing and reporting test results, for example by performing a statistical item analysis to estimate random error effects on scores and to determine if items are functioning properly.
6. Establishing a passing score on which to base a licensing decision.
7. Equating test scores across different test forms, thereby ensuring that scores are comparable whenever multiple exam forms are used (Impara, 1995).

This approach to certification test development was discussed extensively at U. S. Environmental Protection Agency-sponsored test construction workshops held in Kansas City, MO in 1999 and Albuquerque, NM in 2000 (Jeanne Heying, personal communication, 2000).

Historically, these activities have been viewed as comprising a content validation strategy. This is still a very common perspective, but the terminology is potentially confusing. Content validity claims actually tell us more about test subject matter than about what scores mean (Geisinger, 1992). Reference, in this case, to a “content valid test” implies essentially that content and standards have met subject matter expert approval, and not that score interpretation and use, are inevitably

valid. A so-called "valid test" might still yield misleading information or information that gets used in misinformed ways (Frisbie, 2005). Although the component activities of a content validation strategy are necessary, they are not sufficient for making a valid score interpretation (Shepard, 1993). Failure to recognize validity as something more comprehensive than the outcome of a methodological process can lead to claims by a regulatory agency that they are engaged in valid score interpretation and use when they actually may be making little more than a flat assertion about how the test was developed.

In fact, validation is not a process. It is a practical argument (Kane, 1992). The general argument works this way for certification tests: We specify our proposed score interpretation and then implement appropriate test development practices such as job analysis, test plan, and item writing to support a series of inferences from test scores back to job analysis. If each inference is supported by suitable evidence, and if all of our underlying assumptions withstand scrutiny, then we've substantiated our interpretation that test scores indicate achievement in important job-related knowledge and skills necessary for entry-level practice. Note that validation-as-argument is not a proof. There always exist other plausible interpretations against which we must further critically evaluate our proposed score interpretation (Kane, 2004).

Test Score Use

Score interpretation is inextricably linked to score use. Obviously, we are going to base licensing decisions on a passing score, but do scores inform us about what a person, upon receipt of a license, can or will do once on the job? Our proposed score interpretation doesn't speak to that. The scores only tell us about the job knowledge and skill level of test takers who pass the exam. And they tell us almost nothing about people who fail (Messick, 1988).

We assume that individuals who pass a certification exam do so because they possess a sufficient command of job knowledge and skills. The only possible alternatives are that people either cheated or guessed their way to a passing score. But these alternatives are implausible because, in part, of our test construction practices. Consistent exam administration procedures and security measures minimize the possibility of cheating. Random guessing isn't a common test taking strategy, and the laws of probability tell us that correctly guessing even a handful of multiple-choice test items is highly unlikely (Haladyna, 2004). The assumption stands. So, what can we say about people who fail? The obvious assumption here is that they fail because they do not have a sufficient command of job knowledge and skills. In this case, however, there are numerous alternatives that cannot be discounted. People also fail because of inattention, lack of motivation, test anxiety, learning disabilities, reading comprehension problems, language barriers, and other factors, and we don't have any evidence to refute any of these alternatives (Messick, 1988). We cannot say with confidence exactly

why a test taker failed the exam. What can we say about them? Only that they did not pass and will have to take the test again.

Given that we know why people pass but not why they fail leaves us with, at best, a modest assumption about readiness to practice: Persons who pass the exam are more likely to effectively perform entry-level work than those who do not pass. This may strike many as an underwhelming claim, one that instills little confidence in our current testing programs' ability to protect the public welfare. Measurement cannot address that issue, but policy can. In any event, the caution is clear. Be careful not to ascribe more meaning to test scores than they can support.

Potential Pitfalls to Avoid

It is tempting to claim that certification test scores predict future performance. After all, we assume that persons who pass the exam are more likely to effectively perform entry-level work than those who do not pass. Isn't that a type of prediction? Broadly speaking, yes. Intuition permits us to act in the face of uncertainty with the expectation that effective, entry-level practice by persons who pass isn't simply random. However, this is a prediction in the absence of some important evidence. Missing is a line of evidence in the form of a test-criterion correlation (Cronbach, 1980). A criterion is an indicator or marker of effective work performance, such as supervisors' ratings. A correlation is a statistical analysis that measures the strength of the relationship between the test and the criterion. For example, we might establish that higher test scores correlate with more positive supervisor ratings. Leaving aside questions about appropriate criteria and a host of other technical problems, this kind of analysis warrants an evidence-based predictive claim. But our test construction activities generally do not develop this type of evidence. We build certification exams according to a score interpretation that turns on a test's relevance to, and representativeness of, specified content (Guion, 1977). We never get any further than this. Since we lack hard evidence on which to make predictions based on test scores, we should avoid the temptation.

No mention has been made in this discussion about "competence" as a qualifying term associated with certification test scores. Documented mastery of job knowledge and skills is an obviously critical aspect of any licensing decision, but interpreting test scores in terms of an individual's level of competence is a risky proposition. Competence and incompetence are value-laden words that invite a host of unwarranted judgments about the individuals to whom they are applied.

It is still very common to refer to certification tests as competency exams (Schmitt, 1995). The basis for this is the practice of calling job knowledge and skills, collectively, competencies (Williams & Crafts, 1997). Hence, job knowledge and skills tests are competency exams, which encourages us to identify persons who pass them as competent and persons who fail to pass as not competent. This practice

is fine, if all parties understand that competence in this context is narrowly defined as demonstrating qualifying job knowledge and skills. The problem lies in everyday conversation, where the term competence can assume a much more sweeping meaning. Competence, in conventional use, includes other qualities such as interpersonal skills, physical abilities, business acumen, and ethical behavior. As with prediction, our test score interpretation does not speak to these qualities. Referring to competence in regard to score meaning can lead to inappropriate and unfortunate conclusions about test-takers and is best not done, especially when speaking to stakeholders who are likely unfamiliar with the language of assessment.

Conclusion

Modern validity theory holds that validity is a judgment about how well evidence and theory support test score interpretation and use. It is not a test property. Consequently, validation is correctly understood as a practical argument and not a test development process. The argument involves marshalling appropriate evidence in a compelling manner in order to support a proposed score interpretation and to defend it against other plausible interpretations.

A typical score interpretation for pesticide applicator certification exams will probably invoke important job knowledge and skills necessary for effective, entry-level practice. Our exam development activities should generate at least the preliminary evidence to support this interpretation by establishing test content that is relevant to and representative of the knowledge and skills required for entry-level practice. We can then assign limited meaning about knowledge and skills to the scores of individuals who pass the test. This leads us to a plausible assumption that persons who pass are more likely to perform effectively when entering the profession than are persons who fail. Any score-based claim stronger than this is misleading. Certainly, special care should be taken to avoid leading stakeholders to believe that test scores are predictive. Caution must also be exercised when discussing scores in terms of competence. Ascribing more meaning to test scores than they can support does a disservice to test-takers and other stakeholders, and is damaging to the credibility of the credentialing program that allows it.

References

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

Angoff, W. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.). *Test Validity* (pp. 19-32). Hillsdale, NJ: Earlbaum.

Cronbach, L. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement: Measuring Achievement Over a Decade*. In Proceedings of the 1979 ETS

Invitational Conference (pp. 99-108). San Francisco: Jossey-Bass.

Frisbie, D. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, Fall, 21-28.

Geisinger, K. (1992). The metamorphosis of test validation. *Educational Psychologist*, 27 (2), 197-222.

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

Haladyna, T. (2004). *Developing and Validating Multiple-Choice Test Items*. Third Edition. Mahwah, NJ: Earlbaum.

Impara, J. (1995). Overview of the procedures for developing a licensure examination. In Impara, J. (Ed.). *Licensure Testing: Purposes, Procedures, and Practices* (pp. 89-91). University of Nebraska-Lincoln: Buros Institute of Mental Measurement.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 113 (3), 527-535.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, 2 (3), 135-170.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.). *Test Validity* (pp. 33-45). Hillsdale, NJ: Earlbaum.

Messick, S. (1989). Validity. In. R. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: American Council on Education and MacMillan.

Schmitt, K. (1995). What is licensure? In Impara, J. (Ed.). *Licensure Testing: Purposes, Procedures, and Practices* (pp. 3-32). University of Nebraska-Lincoln: Buros Institute of Mental Measurement.

Shepard, L. (1993). Evaluating test validity. *Journal of Research in Education*, 19, 405-450.

Williams, M. & Crafts, L. (1997). Inductive job analysis: The job/task inventory method. In Whetzel, D. & Wheaton, G. (Eds.). *Applied Measurement Methods in Industrial Psychology* (pp. 51-88). Palo Alto, CA: Davies-Black.