ACT Research Report Series 90-3

Generalizing Criterion-related Validity Evidence for Certification Requirements Across Situations and Specialty Areas

Michael T. Kane

May 1990

ACT

Generalizing Criterion-related Validity Evidence for Certification Requirements Across Situations and Specialty Areas

Michael T. Kane

American College Testing

ABSTRACT

Developing good criterion measures of professional performance is difficult. If criterion-related validity evidence for certification requirements could not be generalized beyond the specific context in which it was obtained, gathering that evidence would probably not be worth the effort. This paper examines two possible approaches to the generalization of criterion-related evidence for certification requirements. The first, validity generalization, provides a statistical technique for generalizing the results of particular studies. The criterion problem remains, however; generalizing fluff (criterion-related evidence based on weak and/or inappropriate criteria) merely gives us a more general kind of fluff. The second approach uses substantive models as the basis for generalizing validity data; this second approach offers several advantages, including more emphasis on the nature of the criterion and, possibly, some help in developing better criteria.

ACKNOWLEDGEMENT

I want to thank Pat Benefiel for editing this document and improving both its logic and its style.

Certification provides recognition to individuals who have met some standard of quality in a profession. Generally, applicants for certification must pass a test and meet certain educational, training, and/or experience requirements to demonstrate their proficiency (Shimberg, 1981). Given the purpose of certification, criterion-related validity evidence connecting certification requirements to some criterion of professional performance might seem particularly relevant to the evaluation of various certification requirements.

Part of the appeal of criterion-related validity evidence is that it seems to offer a simple, definitive answer to questions of validity, a "gold standard": unfortunately the simplicity and definiteness dissolve if the criterion is carefully examined. Cronbach (1971) has stated the basic dilemma of criterion-related validity evidence:

> There is a paradox here. The machinery of validation rests on acceptance of the criterion measure as being perfectly valid (save for random error), yet common sense tells one that it is not. . . . Every report of validation against a criterion has to be thought of as carrying the warning clause, "Insofar as the criterion is truly representative of the outcome we wish to maximize,. . ."(pp. 487-488)

Small sample size, restriction of range, and lack of experimental control also present problems in developing criterion-related validity evidence for certification examinations, but the criterion is the main challenge (Shimberg, 1981, 1982; Kane, 1982, 1987).

Criterion-related evidence can be useful, however, if we keep Cronbach's "warning clause" and these additional caveats, in mind. Well designed criterion studies can tell us about the relationships between certification requirements and other variables of interest and can, therefore, be helpful in validating certification requirements.

The issue, then, is how to make the best use of that rare commodity, good criterion-related evidence for certification requirements. This paper focuses on three questions about the generalization of criterion-related validity evidence. First, can meta-analysis techniques, particularly validity generalization (Schmidt & Hunter, 1977), be used to generalize the results of criterion validity studies for certification requirements across situations and specialty areas; and if so, what criteria might be used in deciding whether generalization would be appropriate in a specific case? Second, to what extent can substantive models relating specific certification requirements to specific job activities be used as the basis for generalizing criterion-related data? Third, which of these two approaches, meta-analysis or substantive models, is likely to be more effective in generalizing criterion-related evidence across situations, specialty areas, and specific certification requirements?

Meta-analysis: Validity Generalization

Meta-analysis techniques are designed to draw general conclusions from the results of multiple, independent studies. In examining the criterion validity of certification requirements, the predictor variable can be defined in terms of the binary variable, certified/noncertified, or in terms of achievement on specific requirements (e.g., test scores, level of education, years of experience). The most appropriate meta-analysis technique for analyzing the results of criterion-related studies will depend, in part, on the statistical properties of the predictor variables and the criterion used in the studies. If the predictor is a binary variable and the criterion variable is a continuous variable, methods that cumulate effect sizes across studies (Hedges & Olkin, 1985; Raymond, 1988) would be appropriate. If both

the predictor and the criterion are continuous variables, validity generalization could be used (Schmidt and Hunter, 1977; Hartigan & Wigdor, 1989; Schmidt, Hunter, Pearlman, & Hirsh, 1985; and Sackett, Schmitt, Tenopyr, & Kehoe, 1985).

The data from some early criterion-related studies of employment tests were interpreted as indicating that criterion-related validity coefficients could not be dependably generalized, because these coefficients varied substantially, even across similar jobs in similar or the same settings. This variability was attributed to differences in the requirements associated with specific jobs in different settings and to differences in the settings across organizations and even within an organization (Ghiselli, 1966, 1973). The apparent situational specificity of criterion-related data for employment tests led to suggestions that employment tests be validated separately in every situation in which they were to be used.

In the seventies, Schmidt and Hunter (1977) examined the variability in criterion validity coefficients for employment tests and found that much of this variability could be attributed to such statistical artifacts as sampling error, differences in criterion and test reliabilities, and differences in range restrictions. The results of several studies by Schmidt, Hunter, and their colleagues indicated that the variability in the validity of employment tests across work settings, specific job requirements, etc. is much smaller than had previously been thought and undermined the argument that criterion-related validity coefficients for employment tests are always highly situation specific and cannot be generalized at all. Note, however, that the refutation of an argument (i.e., that a criterion validity evidence cannot be generalized), even if it is completely successful, does not prove the opposite of the argument (i.e., that the evidence can be generalized).

Even though the research on validity generalization indicates that much of the variability in criterion validity coefficients is attributable to statistical artifacts, there is some variability that cannot be explained in this way (Messick, 1989, p. 83; Hartigan and Wigdor, 1989, pp. 130-131). The importance of this residual variability in decision-making has not been determined. To the extent that the residual unexplained variability reflects important differences, validity generalization would not be justified.

A challenge, however, more fundamental than the problems inherent in generalizing criterion-related evidence, is the development of good criterion-related validity evidence in any particular situation. Before we can do much with generalization, we have to get criterion-related evidence, positive or negative, that is worth generalizing.

The major limitation in research on validity generalization is that the original studies employed measures of doubtful validity. It is very difficult to get good measures of performance on the job. Consequently, most studies of the criterion validity of employment tests use supervisors' ratings of performance as the criterion. The validity of such ratings is suspect. Cronbach (1971), describes a study, by Hemphill (1963), of engineers working in industry.

> In a competent study lasting eight years, the Educational Testing Service collected data on thousands of engineers in industry (Hemphill, 1963). Its best testing program could predict only 5 percent of the variance among supervisor's ratings of young engineers. The ratings, it was found, had little to do with technical skill or knowledge and a lot to do with the personal relation that developed between engineer and supervisor. Consequently, the team wound up its criterion-oriented study with the conclusion that the ability tests should play an important part in the company's selection program <u>just because</u> they do not agree with the rating criterion! (Cronbach, 1971, p. 487)

Ratings of highly complex performances made by individual raters (rather than groups of raters), under poorly controlled circumstances, are vulnerable to many sources of bias.

In criticizing some of the conclusions that have been drawn from validity generalization analyses (e.g., that, for a wide range of jobs and settings, performance can be predicted on the basis of measures of general cognitive ability), Prediger (1989) reviewed the results from 34 studies on the relationship between supervisors' ratings and more direct measures of job performance. The data discussed by Prediger were previously summarized by Hunter (1983, 1986) and Heneman (1986). It appears from these data that less than a quarter of the variance in supervisors' ratings is accounted for by differences on performance measures (Prediger, 1989). Earlier, Heneman (1986, p. 818) suggested that, "ratings and results cannot be treated as substitutes for one another." Very little research has been done on validity generalization for criteria other than supervisors' ratings.

The problem with validity generalization, then, is not so much in the "generalization," as it is in the "validity" that is being generalized: the criteria are of doubtful worth. However, Ghisseli's conclusion (1966, 1973) that validity coefficients are highly situation specific rested on these same types of data. By showing that the existing evidence for a very high degree of situational specificity largely evaporates when statistical artifacts are carefully considered, Schmidt and Hunter effectively challenged the view that the criterion-related validity evidence of employment tests should never be generalized.

Some of the claims that have been made for validity generalization are disputed (e.g., Hartigan and Wigdor, 1989; Prediger, 1989; Sackett et. al., 1985), but the work of Schmidt and Hunter and their associates (Schmidt and Hunter, 1977; Schmidt et. al., 1985) has provided support for generalizing criterion validity evidence in cases where aptitude tests are used to predict supervisors' ratings of job performance. Note that the validity

generalization results do not correct the potential problems in studies using supervisors' ratings as criterion measures, but they could support inferences from previous studies using these criteria to new studies using similar criteria.

Most research on validity generalization has been based on studies of employment testing in business, industry, and government in which aptitude tests are used to predict supervisors' ratings of performance. Certification requirements are usually designed as measures of knowledge and skill rather than as aptitude measures of the kind used in employment testing. The criteria that would be of most interest in evaluating certification requirements would be based on the direct assessment of professional performance or on client outcomes. Since validity generalization analyses have not been applied to the measures, criteria, and situations that are most relevant to certification programs, these analyses do not justify the generalization of criterion-related validity evidence for certification requirements to new situations or specialty areas.

Guidelines for Generalizing Criterion-related Validity Evidence (Using Meta-Analysis)

The research that has been done on validity generalization does not provide direct support for generalizing criterion-related evidence for certification requirements. The old, absolute prohibition against generalizing validity evidence may have been undermined, but the evidence supporting generalization is largely restricted to jobs in business and industry, and suffers from the use of relatively weak criterion measures in the underlying studies.

However, it may be possible to apply these meta-analysis techniques to certification requirements in the future. The discussion of validity generalization in the current version of the Standards for Educational and Psychological Testing (AERA, APA, NCME 1985) contains a limited and cautious endorsement of validity generalization:

> In conducting studies of the generalizability of validity evidence, the prior studies that are included may vary according to several situational facets. Some of the major facets are (a) differences in the way the predictor construct is measured, (b) the type of job or curriculum involved, (c) the type of criterion measure, (d) the type of test takers, and (e) the time period in which the study was conducted. In any particular study of validity generalization, any number of these facets might vary, and a major objective of the study is to determine whether variation in these facets affects the generalizability of validity evidence.
>
> The extent to which predictive or concurrent evidence of validity generalization can be used as criterion-related evidence in new situations is in large measure a function of accumulated research. Consequently, although evidence of generalization can often be used to support a claim of validity in a new situation, the extent to which this claim is justified is constrained by the available data. (p. 12)

To summarize: Validity generalization can sometimes be used as part of an argument for validity in a new situation, but such claims should be supported by data, and the extent to which the claim is justified is constrained by the available data.

There is one standard, 1.16, on validity generalization (AERA, APA, NCME, 1985):

> Standard 1.16 When adequate local validation evidence is not available, criterion-related evidence of validity for a specified test use may be based on validity generalization from a set of prior studies, provided that the specific test-use situation can be considered to have been drawn from the same population of situations on which validity generalization was conducted. (Primary)
>
> **Comment:**
> Several methods of validity generalization and simultaneous estimation have proven useful. In all methods, the integrity of the inference depends on the degree of similarity between the local situation and the prior set of situations. Present and prior situations can be judged to be similar, for example, according to factors such as the characteristics of the people and job functions involved. Relational measures

> (correlations, regressions, success rates, etc.) should be carefully selected to be appropriate for the inference to be made. (pp. 16-17)

Standard 1.16 and the previous discussion suggest at least two criteria for evaluating the appropriateness of a particular application of validity generalization.

First, the new situation (setting, specific job, specific test) to which criterion-related evidence is to be generalized or extrapolated should be similar to the situations for which criterion-related data are available. For certification requirements, the results of a job analysis study providing data on work settings, specialty areas, etc. could be used to examine the question of similarity. For example, A Study of Nursing Practice and Role Delineation and Job Analysis of Entry-level Performance of Registered Nurses (Kane, Kingsbury, Colton, and Estes, 1986), published by the National Council of State Boards of Nursing, Inc., provides detailed information about practice patterns in nursing across work settings, shifts, areas of practice, etc. In the absence of empirical job analysis data, expert judgments of the characteristics of the work settings, job responsibilities, etc. in the new situation could be employed.

Second, the inference to a new situation should generally be based on criterion-related data from several studies rather than on data from just one study. If a cluster of studies involving different settings, job responsibilities, and/or criterion measures yield similar results, we can have confidence that a relationship of some generality exists. The pattern of relationships within the cluster supports the generalization of the findings to a new situation that is similar to those in the cluster of existing studies.

Generalization from a single study to a new situation is much more hazardous, and should be labeled "extrapolation" rather than "generalization." The results of any single study may be strongly influenced by characteristics

specific to the situation studied, and these characteristics may not have much impact in most of the situations over which generalization is intended.

Meta-analysis methods, including validity generalization, are most useful when applied to a large number of studies. Because of the sampling variance over studies, the results of the meta-analysis are not likely to be accurate if they are based on only a few studies. Because good criterion-related studies are difficult to conduct for certification requirements, it is not likely that the results of a large number of such studies will be available in the foreseeable future. Therefore, meta-analysis is not likely to yield very dependable results for studies involving certification requirements.

In addition, there is a potentially serious risk associated with reliance on validity generalization. This technique summarizes the results of separate criterion-related studies. While it demands results from many studies, it largely ignores most of the details of the individual studies. This is not necessarily a bad thing; every method for summarizing data highlights some differences and supresses others. However, if we combine validity generalization's need for results from a large number of studies and its lack of concern about the quality of individual studies with the difficulties inherent in developing and implementing good criteria, it seems likely that, to the extent that we rely on validity generalization, we will probably give very little attention to developing good criterion measures. Validity generalization may be useful in summarizing criterion-related validity evidence, but it may also encourage the operation of a form of Gresham's Law for criterion-related evidence.

Using Substantive Models to Generalize Criterion-Related Validity Evidence

The direct generalization of criterion-related validity evidence can be appropriate and useful under some circumstances. However, because it suggests a fairly straightforward statistical inference, meta-analysis or, more particularly, validity generalization, does not provide a comprehensive framework for the use of criterion-related data in validating certification requirements. It oversimplifies the situation and focuses on a narrow range of data when many different types of data are relevant. Using empirical data to test theoretical models provides a broader and more realistic framework for drawing general conclusions from the results of specific criterion-related validity studies, an approach in which generalization plays an important, but subordinate role.

The general structure of the interpretive model implicit in most certification requirements is not very esoteric or technical and depends more on common sense than on any deep analysis of the nature of things, but like any scientific theory, the model includes assumptions and makes predictions. Cronbach (1989) has recently suggested that the models used in test score interpretations could be referred to as "constructions," to reflect their status as loose collections of assumptions, partial theories, etc.

The model, or "construction," connecting certification requirements to job-related performance has the following general form for most certification programs:

> Premise 1: There are certain domains of knowledge and skill that are relevant to the quality of practice, in the sense that practitioners with a high level of achievement in these domains are prepared to provide professional services of high quality to their clients.
>
> Premise 2: Meeting the certification requirements indicates a high level of achievement in the appropriate domains of knowledge and skills.

> Conclusion: Meeting the certification requirements indicates that practitioners are prepared to provide professional services of high quality to their clients.

The argument is deceptively simple; the difficulties are hidden in the premises.

The details of the model tend to get complicated: the content of the domains may be very extensive and complex; defining the limits and structure of the domains is difficult; setting standards of achievement in the domains raises both technical and sociopolitical issues. Nevertheless, the basic model relating certification requirements to professional performance practice is simple. It assumes that the possession of certain knowledges and skills can improve professional performance and that the certification requirements provide indicators of these knowledges and skills, thus implying that certification should be related to professional performance.

We can check on the plausibility of the model in a specific case by spelling out the various assumptions incorporated in the two premises and then testing these assumptions against experience. For example, the first premise states that there are knowledge and skill domains that can be shown to be relevant to professional performance. In fleshing out this premise for a specific certification program, we would specify the content of these domains, their boundaries, and their internal structure. The argument that a certain knowledge or skill in one of these domains is related to professional performance is based on two further assumptions: (1) that the knowledge/skill is necessary in order to perform certain activities, and (2) that these activities occur in practice. The first of these two assumptions tends to be justified by logical analysis (suggesting, for example, that knowledge of natural childbirth techniques is helpful if one is going to teach classes on natural childbirth) and/or clinical research (showing, for instance, that

specific natural childbirth methods are, in fact, effective). The assumption that specific groups of practitioners perform the activities (e.g., teaching childbirth classes) can be checked in an empirical job analysis. That is, Premise 1 is supported by showing that the domains are related to important parts of practice.

Premise 2, which relates the certification requirements to the knowledge and skill domains, can be checked in many ways. For example, test items can be subjected to criticism from content experts and experienced item developers (Cronbach, 1971, p. 457); the processes that candidates use in responding to test questions can be examined by having candidates work through questions aloud or by asking candidates why they responded in various ways (Cronbach 1971, p. 474). Techniques for evaluating reliability (Feldt & Brennan, 1989) and for identifying possible sources of bias in the requirements (Cole & Moss, 1989) have been developed. In some cases it may be particularly useful to examine the relationship between specific components of the certification requirements and the results of using other methods to measure the domains of knowledge and skill (Campbell & Fiske, 1959). Premise 2 is supported by showing that the relationship between the certification requirements and achievement in the appropriate domains of knowledge and skill is plausible and can stand up to serious criticism.

A series of empirical tests of specific assumptions can generate high confidence in the model describing the relationship between certification requirements and professional performance, especially if the model implicit in these requirements is reasonable to begin with. None of these analyses involve the comparison of certification requirements to a criterion measure, but taken as a whole, they can provide a strong case for the claim that the certification requirements reflect the appropriate domains of knowledge and skills.

We may also be able to check on the plausibility of the model by collecting criterion-related evidence. A positive relationship between certification and professional performance provides support for the model, and thus for the appropriateness of the certification requirements. A failure to find a positive relationship casts doubt on the model. However, data describing the relationship between certification (or specific certification requirements) and a criterion measure of professional performance does not provide a completely decisive test of the model in either of these two cases; the criterion is never beyond question, the control of extraneous factors is never complete, and sample sizes are usually too small to effectively control sampling errors. The observed relationship between the certification requirements and the criterion can have a strong impact on our faith in the model, but it is usually not decisive.

The criterion-related evidence is one of several types of evidence with which the general model which can be tested. Criterion-related studies have the advantage of allowing us to test the model as a whole, rather than its parts, but they also involve many problems of interpretation, including, in most cases, a questionable criterion and small sample sizes. There is no compelling reason to give criterion-related evidence a privileged position in evaluating the validity of certification requirements.

Nevertheless, criterion-related evidence does provide one way to test the model, and this evidence can be generalized through its impact on the credibility of the model. To the extent that criterion-related evidence supports a general model, it provides support for the validity of the certification requirements in any context in which the relationships specified in the model are expected to apply. Since the model applies to a variety of settings and criterion measures,

it provides a basis for generalizing or extrapolating current experience to new situations.

Models achieve their power as generalizers of empirical evidence, in part, by providing an explanation of the observed relationship. If it is reasonable to assume that certification requirements reflect certain knowledge and skills, that the knowledge and skills are necessary for the successful performance of activities constituting an important part of practice, and that the criterion measure depends to a substantial extent on the performance of the activities requiring these skills, then it is reasonable to expect a positive relationship between certification and performance in those situations where the model can be shown to apply. The model provides a basis for generalizing criterion-related evidence to similar situations and a basis for identifying the situations that are "similar."

The assumptions in the model underlying certification can facilitate the generalization of criterion-related evidence. The model can help to define reasonable bounds for generalization. It can indicate cases in which generalization is not reasonable. It can help us to design our studies and develop criteria. The model is essential for thoughtful generalization and reasonable extrapolation.

Substantive Models vs. Meta-analysis

There are at least four reasons for thinking that substantive models would be preferrable to statistical meta-analysis models as a foundation for generalizing criterion-related validity evidence for certification requirements.

First, meta-analysis is designed to extract useful, general conclusions from large numbers of empirical studies that do not provide entirely consistent results. If the studies yielded highly consistent results, we

would probably not need a meta-analysis to identify these results. Meta-analysis tends to focus on very general properties of the studies, particularly estimates of effect sizes and study characteristics, like sample size, that might influence effect sizes. These analyses tend to ignore many of the details of the studies (e.g., the relationship between the content of the test and the content of the criterion measure in a criterion-related validity study). As a result, meta-analysis is likely to be most useful when a large number of studies are relevant to the issue under consideration and these studies can be considered more-or-less interchangeable. This is not the case in examining the validity of certification requirements. As noted throughout this paper, we are not overwhelmed by large numbers of criterion validity studies that employed strong criteria, and we are not likely to have the luxury of this problem in the near future.

Second, the substantive model provides a basis for defining the limits of generalization. The substantive model assumes that certain activities occur in practice, and, therefore, the model would apply to situations in which the activities occur and would not apply to situations in which the activities do not occur.

Generalization is always based on judgments about similarity and differences (Cronbach, 1971, p.485), and a model or theory of the relationship between certification requirements and performance plays a major role in determining the situations, specialty areas, and criteria that can be considered similar. The model can provide a powerful mechanism for generalizing the results of even a few studies to a wide range of situations and can indicate the range of situations, criterion measures, etc., to which the results can be generalized.

The model may also identify cases where generalization is not appropriate. For example, even if we have a number of studies in which certification requirements emphasizing technical skills are strongly related to criteria emphasizing technical proficiency, we would probably not be justified in extrapolating to a situation which involved a criterion measure of client satisfaction. The model can, therefore, help us to make sense out of conflicting results by allowing us to distinguish between situations where the model should apply and situations where it should not apply.

Third, the substantive model gives us a basis for evaluating the appropriateness of a criterion measure. A criterion measure that focused on the activities emphasized within the certification model could provide useful information about the model. A criterion measure that did not focus on the activities emphasized within the model would not provide a good basis for evaluating the model. A conflict between the activities included in the model and those included in the criterion suggests that either the model or the criterion is mispecified; one or the other (or perhaps both) does not adequately reflect practice requirements. By clarifying the relationships among knowledge and skill domains, specific activities performed in practice, and the quality of practice, the model can help us to develop criteria that reflect the relevant outcomes. In selecting criteria, it is important to consider the possible side effects of certification as well as the intended outcomes (Dvorak, 1990).

Fourth, the substantive model gives us a basis for recognizing factors specific to a given situation, that might influence the outcomes of a criterion-related validity study. For example, it is conceivable that, in a certain work setting, the staff who are not certified have levels of education and experience that are similar to those who are certified.

As noted earlier, generalization always relies on some assumptions, implicit or explicit, that provide a basis for identifying situations, criteria, etc. that are similar or different. A major concern in evaluating certification requirements against criteria of performance in practice is our inability to control what goes on in practice, and the model may help us to identify cases where this lack of control is especially problematic. For example, it may happen that most of the uncertified practitioners in an area have met most of the requirements for certification, such as education and experience; these practitioners may simply have not bothered to become certified, perhaps because they dislike paperwork or tests. In this situation, the logic of certification would not lead us to expect a relationship between certification and performance.

Conclusions

The argument in this paper can be summarized in terms of four major points:

First, the work of Schmidt and Hunter, and others, on validity generalization has undermined the older, blanket prohibition against generalizing criterion-related validity evidence. The researcher seeking to generalize findings from a specific study or a small set of studies to more general conclusions about the relationship between certification requirements and criterion measures does not have to overcome a mass of evidence against validity generalization. Schmidt and Hunter have done that.

Second, the work on validity generalization has not shown that validity evidence relating certification requirements to performance in practice can be generalized to new situations, specialty areas, and specific requirements. The studies analyzed by Schmidt and Hunter generally involved the relationship

between aptitude-based employment tests and supervisors' ratings. These results do not directly support validity generalization for certification requirements. Although research on validity generalization has cleared the air in some ways, the researcher seeking to generalize criterion-related validity evidence for certification requirements still needs to make the positive case for generalization.

Third, the "criterion problem" is still the major stumbling block. One must get good criterion-related evidence before one can achieve much by generalizing this evidence. The work on validity generalization does nothing to solve the criterion problem.

Fourth, an alternate, and perhaps more effective, route to the generalization of criterion-related validity evidence for certification is to treat such evidence as providing an empirical check on the model relating certification requirements to performance in practice. To the extent that we understand the relationship between certification requirements and the criterion in specific situations, we tend to have more faith in the relationship, and we have a basis for generalizing the results to "similar" situations. The model also provides guidance for deciding on the range of new situations to which generalization would be legitimate, and for designing new studies and developing criterion measures.

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.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56 81-105.

Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.) Educational Measurement (Third Edition) New York: American Council on Education and Macmillan.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.) Educational Measurement (Second Edition). Washington, DC: American Council on Education.

Cronbach, L. (1989). Construct validation after thirty years. In R. L. Linn (Ed.) Intelligence: Measurement, theory and public policy. Urbana: University of Illinois Press.

Dvorak, E. (1990). Relationship between certification and practice outcomes. Paper presented at annual meeting of American Educational Research Association, Boston, 1990.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.) Educational Measurement (Third Edition) New York: American Council on Education and Macmillan

Ghiselli, E. E. (1966). The validity of occupational aptitude tests. New York, NY: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Hartigan, J. A. & Wigdor, A. K. (1989). Fairness in employment testing: Validity generalization, minority issues, and the general aptitude battery. National Academy Press: Washington, DC.

Hedges, L. & Olkin, I. (1985). Statistical Methods for Meta-analysis. New York: Academic Press.

Heneman, R. L. (1986). The relationship between supervisory ratings and result-oriented measures of performance: A meta-analysis. Personnel Psychology, 39, 811-826.

Hemphill, J. K. (1963). The Engineering Study. Englewood Cliffs, NJ: Prentice Hall.

Hunter, J. E. (1983). A casual analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory (pp. 257-266). Hillsdale, NJ: Erlbaum.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. Journal of Vocational Behavior, 29, 340-362.

Kane, M. (1982). The validity of licensure examinations. American Psychologist, 37, 911-918.

Kane, M. (1986). The future of testing for licensure and certification examinations. In B. Plake and J. Witt (Eds.), The Future of Testing and Measurement, Hillsdale, NJ: Erlbaum, 145-181.

Kane, M. (1987). Is predictive validity the gold standard or is it the holy grail of examinations in the professions? Professions Education Research Notes, 9, 9-13.

Kane, M., Kingsbury, C., Colton, D. and Estes, C. (1986). A Study of Nursing Practice and Role Delineation and Job Analysis of Entry-level Performance of Registered Nurses, Chicago, IL: National Council of State Boards of Nursing, Inc.

Prediger, D. J. (1989). Ability Differences across Occupations: More than g. Journal of Vocational Behavior, 34, 1-27.

Raju, N. S., Pappas, S., & Williams, C. P. (1989). An empirical Monte Carlo test of the accuracy of the correlation, covariance, and regression slope models for assessing validity generalization. Journal of Applied Psychology, 74, 901-911.

Raymond, M. (1988). The relationship between educational preparation and performance on nursing certification examinations. Journal of Nursing Education, 27, 6-9.

Sackett, P. R., Schmitt, N., Tenopyr, M. L., & Kehoe, J. (1985). Commentary on forty questions about validity generalization and meta-analysis. Personnel Psychology, 38, 697-798.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F. L., Hunter, J. E., Pearlman, K. & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. Personnel Psychology, 38, 697-798.

Shimberg, B. (1981). Testing for licensure and certification. American Psychologist, 36, 1138-1146.

Shimberg, B. (1982). Occupational licensing: A public perspective. Princeton, NJ: Educational Testing Service.