ED 476 421 TM 034 923

AUTHOR Kane, Michael; Case, Susan

TITLE The Reliability and Validity of Weighted Composite Scores.

PUB DATE 2003-04-00

NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Reliability; Simulation; *Validity; *Weighted Scores

IDENTIFIERS *Composite Scores

ABSTRACT

The scores on two distinct tests (e.g., essay and objective) are often combined into a composite score, which is used to make decisions. The validity of the observed composite can sometimes be evaluated relative to a separate criterion. In cases where no criterion is available, the observed composite has generally been evaluated in terms of its reliability. The simulation analyses in this study are based on a simple content-based model for the validity of the observed composite as an estimate of a target composite, based on a priori weights for the target composite and the universe scores (or true scores) for the two tests. The results suggest that giving extra weight to the more reliable of the two observed scores tends to improve the reliability of the composite, and, up to a point, tends to improve its validity. Giving too much weight to the more reliable scores can decrease the validity of the observed composite as a measure of the target composite. (Contains 2 figures and 13 references.) (SLD)

Combine scores M02.wpd

The Reliability and Validity of Weighted Composite Scores

Michael Kane
Susan Case

Paper presented at Annual Meeting of NCME, Chicago, 2003

Abstract

The scores on two distinct tests (e.g., essay and objective) are often combined into a composite score, which is used to make decisions. The validity of the observed composite can sometimes be evaluated relative to a separate criterion. In cases where no criterion is available, the observed composite has generally been evaluated in terms of its reliability. The analyses in this paper are based on a simple, content-based model for the validity of the observed composite as an estimate of a target composite, based on apriori weights for the target composite and the universe scores (or true scores) for the two tests. The results.suggest that giving extra weight to the more reliable of the two observed scores tends to improve the reliability of the composite, and up to a point tends to improve its validity. Giving too much weight to the more reliable score can decrease the validity of the observed composite as a measure of the target composite.

running head: Validity of composite scores

The Reliability and Validity of Weighted Composite Scores

Michael Kane
Susan Case

In many achievement-testing contexts it is desirable to be able to combine measures of distinct competencies into a single variable. For example, we might want to test the students' breath of knowledge in some domain and their ability to use this knowledge in some way. In testing breadth of knowledge, objective tests consisting of a large number of items drawn from all parts of the domain are likely to be particularly effective. To assess skill in using knowledge (e.g., to analyze a trend or solve a problem) a performance test of some kind generally seems more appropriate. As a result, it is not unusual to design high-stakes tests(e.g., certification tests) with two or more parts (e.g., an objective component and an performance component) and to score these two components separately (Klein, 1995; Mehrens, 1990: Rudner, 2001; Wainer and Thissen, 1993).

There are two general approaches to decision making where the decision is to be based on more than one test score: compensatory and non-compensatory. A non-compensatory decision rule considers the separate component scores independently, and it is necessary to pass each components in order to pass the test as a whole. A familiar example of a non-compensatory system is found in the tests used to award driving licenses. In most jurisdictions, one must pass both an objective test covering the rules of the road and a road test assessing essential driving skills. A high score on one component cannot be used to compensate for a low score on the other.

In a compensatory system, the scores on the separate components are combined into a single overall score, which is then used as the basis for a decision. The most common way to combine the component scores is to take their weighted sum or average. In a compensatory scoring system, a high score on one component can compensate for a low score on another component.

Most high-stakes testing programs that rely on two or more tests to make a decision use compensatory scoring rules. There are at least three reasons for this. First, in cases where the different competencies can compensate for each other in actual practice situations (e.g., the tendency for professionals to specialize in different areas of practice), a compensatory rule makes sense. Second, compensatory scoring systems tend to be much more reliable than non-compensatory rules (Hambleton and Slater, 1997), because longer tests (e.g., the composite) tend to be more reliable than shorter tests (e.g., the subtests). Third, compensatory scoring systems are especially appealing when there is considerable overlap in the knowledge and skills being assessed by the different tests (e.g., on objective and essay tests in the same content area).

A non-compensatory rule tends to make more sense in cases where the skills measured by the different tests are clearly distinct (e.g., written test on rules of the road and a driving test), are each necessary for effective performance, and are each measured with adequate reliability. If any one of these conditions fails, a compensatory rule is recommended.

In a compensatory scoring system, the issue is how to combine the scores from the two components in generating a final score. Assuming that the final score is to be a weighted sum of the separate component scores, the question is reduced to the choice of the weight to be assigned to each component.

Previous Research

Research on this issue goes back a long way (Kelly, 1927). In the absence of an external criterion to be used in determining weights, the focus tended to be on the reliability or the standard errors of the test scores. Gulliksen provides an extended discussion of early research on weighting and differential prediction (Gulliksen, 1950, pp. 312 - 359). In cases where no external criterion is available, Gulliksen (1950) suggests two alternatives: to weight the components so as to maximize the reliability of the composite, or to rely on expert judgement for the determination of the weights (p.357).

Jarjoura and Brennan (1982) examined the problem of estimating the universe score for a test developed to match a table of specifications. They assumed that the categories were fixed and that some number of items would be randomly sampled from each category. Jarjoura and Brennan (1982) took the variable of interest for each person to be a weighted composite of his or her category universe scores, with the category weights pre-specified (explicitly or implicitly in terms of the number of items in each category). Among other things, they examined the possibility of selecting the weights to be assigned in a composite of the observed category scores in order to minimize the error variance in estimating a universe-score composite with fixed apriori weights. After analyzing data from a multiple-choice test with five categories, they concluded that it was, "important to have at least some balance of items across the categories" (Jarjoura and Brennan, 1982), but that, "strict adherence" to the prescribed weights was not critical.

Wainer and Thissen (1993) examined this issue in the context of advanced placement tests, that typically include essay questions and multiple-choice questions. They assumed that the goal of the assessment is to estimate some unidimensional variable. They suggest that "the use of a single summary score carries the clear implication that both parts of the test - multiple choice and constructed response - are presumed to measure a single dimension" (p.104). This conclusion echoes a suggestion made by Gulliksen over 50 years ago:

> Sometimes when a battery of tests is given, the problem is to obtain a single score from the battery. This implies that the battery is a one-factor battery or is to be treated as a one-factor battery. (Gulliksen, 1950,p. 351)

This conclusion is unwarranted. In making high-stakes decisions, it is not uncommon to combine scores purposely designed to measure different dimensions of skill or competence. On the Advanced Placement examinations, on college-admission tests (SAT or ACT), on many state testing programs, and on licensure examinations, scores on distinct subtests are combined to form a composite score, even though the different subtests are designed to measure different sets of competencies.

In all of these cases, the different sets of competencies or dimensions measured by the different component measures are viewed as being distinct and important. They are expected to be positively correlated, but they are not viewed as being exchangeable. Many high-stakes educational tests (e.g., ACT, SAT) include a quantitative subtest and a verbal subtest. The content of these different subtests has essentially no overlap (verbal tasks vs. mathematical problems). These separate tests are not seen as measuring the same set of competencies, although we are not surprised that they tend to be positively correlated. The scores are combined mainly to enhance validity of the total score as an overall measure of readiness for college. The use of a single composite score may also improve reliability. Note that the competencies measured by the verbal and quantitative subtests are likely to be at least partially compensatory; in terms of success in college, strength in one set of competencies can compensate for weakness in another, by choice of major and by time devoted to different tasks.

The model proposed by Wainer and Thissen (1993), which assumes that the subtests all measure the same dimension may be reasonable in some contexts, but it does not adequately represent the relationships among component subtests in many high-stakes testing programs. Wainer and Thissen suggest that the usefulness of a test component depends on its reliability and validity, but since good measures of differential validity are often lacking and reliability puts an upper bound on validity, they opt to "use reliability as a surrogate for test usefulness" (p.109).

> "... it is sensible to either weight the components by some function of their reliability or to modify the lengths of one or more of those components to make them equal in reliability." (Wainer & Thissen, 1993, p.116)

However, given the basic assumption of the Wainer and Thissen (1993) analysis, the second of these two assumptions does not make much sense. If the two subtests are measuring the same dimension, there is no reason not to rely exclusively on the more efficient approaches. Later in their article, Wainer & Thissen (1993) conclude that constructed-response items "take more examinee time and resources to measure essentially the same thing more poorly than" multiple-choice items. (Wainer & Thissen, 1993, p.116). This conclusion is perfectly valid if the various subtests are viewed as different ways of measuring essentially the same thing. It is not valid, if the two subtests are in fact measuring separate components of competence, even if these components are positively correlated.

In their analysis of tests based on tables of specification, which was discussed earlier, Jarjoura and Brennan (1982) reached essentially the opposite conclusion from Wainer and Thissen. They examined how to distribute a fixed number of items over categories in order to minimize the error in estimating the composite universe score. Their results suggest that the optimal distribution of items over categories would tend to assign more items to the categories with the largest error. Even though the average of the universe score correlations among categories was 0.93, with a range of 0.88 to 0.97, Jarjoura and Brennan (1982) did not suggest eliminating some categories in favor of others.

On a practical level, the difference in conclusions between Wainer and Thissen (1993) and Jarjoura and Brennan's (1982) may be attributable to the fact that it is much easier to add multiple-choice items to categories than it is to add essay questions to an examination. However, at a more fundamental level, the difference in conclusions results from their different goals, and therefore their different criteria for success. Wainer and Thissen (1993) set as their goal the enhancement of the overall reliability of the examination, and found (not surprisingly) that the most effective way to achieve this goal was to focus on the most reliable subcomponents, the objective test. Jarjoura and Brennan's (1982) analysis did not seek to maximize reliability, but rather to maximize the agreement between the observed composite score and the pre-specified target composite score (a validity issue), which led to the conclusion that all of the components in the composite should be measured reasonably well.

Rudner (2001) assumed the existence of a criterion variable against which the validity of the observed composite can be evaluated, and then focused on the validity of the observed composite score as a function of its component weights. Using this model, he developed formulas for the reliability and validity of the composite, as a function of the component weights. He also presents a hypothetical example in which validity systematically decreases as the reliability is increased by giving more weight to the more reliable but less valid component at the expense of a less reliable but more valid component. It is not practical to use Rudner's (2001) analysis in many cases, because no external criterion is available, but Rudner does make the case that increasing the reliability of an observed composite is not necessarily an effective way to improve its validity. This paper makes a similar point without assuming the existence of an external criterion.

The Surrogation Fallacy

In this context, it is important to recognize that just as correlation does not imply causality, even a perfect correlation between two measures does not imply that they have the same meaning. Among copper rods of a certain diameter, length will be perfectly correlated with weight, but length is not the same thing as weight; and if we consider a broader universe of objects (e.g. rods of different diameters or different materials) the perfect correlation between length and weight will break down. Similarly, among students who have graduated from American high schools, the correlation between scores on a mathematics test and an English usage test would generally be positive and fairly high; but competence in English is not the same thing as competence in mathematics. The high correlation between English and math scores reflects the fact that successful students in good American high schools typically do well on both English and math tests, and poorer students in poorer schools do less well. There are, of course exceptions (students who are stars in one subject and failures in the other), but these are rare events. However, if we broaden the context, the magnitude of the correlation between English and Math would be likely to decrease substantially. For example, if we dropped math from the standard high-school curriculum or broadened the population of interest to include students from all over the world, we would expect these correlations to drop sharply.

Scriven (1987) addresses this issue in some detail and maintains that the, "use of a correlate....as if it were an explanation of, or a substitute for, or a valid evaluative criterion of, another variable" is an example of what he calls the fallacy of statistical surrogation" (p.11). The fallacy involves a "substitution of a statistical notion for a concept of a more sophisticated kind such as causation or identity." (Scriven, 1987, p.11)

The surrogation fallacy has potentially serious implications for high-stakes testing. If the test has two components, A and B, each of which measures a distinct but overlapping set of competencies, the test takers are likely to prepare for both parts. To the extent that the two content domains overlap, general preparation for A and B might be pretty much the same. However, specific preparation for A is likely to focus on the domain of tasks included in the A component of the test, and specific preparation for B will focus on its task domain.

Now suppose that A and B are highly correlated (i.e., the disattenuated correlations approaches 1.0), and we drop B in favor of A, perhaps because the A format is more convenient or cheaper, or has better technical characteristics (e.g., reliability). What happens? Specific preparation for A is likely to increase and specific preparation for B is likely to decrease. If A is a multiple choice test, and B is an essay test, test preparation will emphasize practice on answering multiple-choice questions rather than on writing essays. If specific test preparation is a major part of total academic preparation (as it can be in high-stakes contexts), this may be seen as a serious loss, especially if writing extended responses is considered to be an important part of the domain of ultimate interest.

The Model

In this paper, we will employ criteria based mainly on validity (and on reliability as a necessary condition for validity) in evaluating various observed composite scores as measures of the target composite score. Validity is generally recognized as a more fundamental and more important criterion than reliability in evaluating measurement procedures. In modeling the validity of the composite, we will assume that the two components are designed to measure separate, but not necessarily independent, aspects of performance, rather than assuming that they measure the same dimension, as suggested by Gulliksen (1950) and Wainer and Thissen (1993).

We will not assume that an external criterion is available for the composite score or for either of the components. Rather, we will assume that the two measures have been developed as samples of distinct kinds of performance, both of which are considered relevant to the decision to be made. The analysis of the validity and reliability of the composite score relies on information about the means, variances, and reliabilities of the two tests, and the correlation between the two tests, rather than on any external criterion. The statistical model employed here is similar in spirit and assumptions to Jarjoura and Brennan (19820, although notation and formulation are different to accomodate a different context. Jarjoura and Brennan (1982) examine the weights of category scores in a MC test; we are interested in combining scores for different test components (e.g., MC and essay).

The Observed Composite Variable

The fact that decision makers go to the trouble and expense of developing and administering two tests (e.g., an objective test and a performance test) indicates that they think that the two tests are measuring different sets of competencies (possible with substantial overlap) and that both sets of competencies are important.

It is reasonable to consider each of the tests as a measure of a domain of knowledge, skill or judgement relevant to the decision to be made. There may be considerable overlap in the two sets of competencies, but the overlap is not expected to be perfect.

The validity evidence for most high-stakes testing programs (e.g., licensure and certification test, high school graduation tests), are based mainly on expert judgements about the appropriateness of the tasks included in the assessment. The variable measured by each component test tends to be defined in terms of the content and form of the assessment and can be viewed as the expected score over the domain of tasks from which those in the assessment are drawn. In classical test theory, this variable would be called the 'true score." In generalizability theory, the domain would be referred to as the universe of generalization, and the expected value over this domain would be referred to as a "universe score."

The two observed variables include sampling error, and can be represented as:

$$X_1 = \mu_1 + e_1$$

and

$$X_2 = \mu_2 + e_2$$

where μ_1 and μ_2 are the true scores or universe scores for the two measures, and e_1 and e_2 represent the randon errors associated with the two measures. The errors are assumed to have means of 0.0 and standard deviations of $\sigma(e_1)$ and $\sigma(e_2)$ and to be uncorrelated with each other, with μ_1 and μ_2, and with all other variables. Therefore, the observed score X_1 has a mean of $E(\mu_1)$ and a variance given by:

$$\sigma^2(X_1) = \sigma^2(\mu_1) + \sigma^2(e_1)$$

Similarly, X_2 has a mean of $E(\mu_2)$ and a variance of:

$$\sigma^2(X_2) = \sigma^2(\mu_2) + \sigma^2(e_2)$$

For convenience, we will assume that the observed scores have been scaled to have equal variances.

The reliabilities of X_1 is given by::

$$\rho_{11} = \frac{\sigma^2(\mu_1)}{\sigma^2(X_1)} \quad \textit{and} \quad \sigma^2(\mu_1) = \rho_{11}\sigma^2(X_1)$$

and the reliability of X_2 is:

$$\rho_{22} = \frac{\sigma^2(\mu_2)}{\sigma^2(X_2)} \quad \textit{and} \quad \sigma^2(\mu_2) = \rho_{22}\sigma^2(X_2)$$

Because the error terms are uncorrelated with each other and with all other variables, the covariance between the two target scores is equal to the covariance between the two observed scores:

$$cov(\mu_1,\mu_2) = cov(X_1,X_2) = cov_{12}$$

The means, variances, and reliabilities of of X_1 and X_2, as well as the covariance between X_1 and X_2, are all observable quantities.

The Target Variable

As noted above, we assume that the two observed scores X_1 and X_2 represent measures of two distinct but not necessarily independent target variables, μ_1 and μ_2.

The goal is to measure two separate components of competence represented by the two component tests and to combine these two scores into a weighted composite. Since we usually do not have criterion measures for either component or the composite, it is reasonable to take their true scores as our best indicators of the variables of interest. What we want is an observed composite score that is a good measure of the target composite score, defined as a weighted average of the two competencies.

Each of the tests is designed to measure the competencies reflected in the tasks included in that test. The target variable for each test can therefore be conceptualized in terms of a universe score or true score associated with that test. This is consistent with how the separate tests are typically developed and used. The items/tasks included in each test are designed to measure a subset of the competencies required in a particular context. For example, the multiple-choice tests used for credentialing programs are designed to measure command of the knowledge base of the profession, including competence in applying this knowledge to simple examples. Essay tests and other performance tests tend to focus on more complex and realistic applications, on problem solving, and perhaps on writing skill. In each case, the test itself may be the best available measure of the specific set of competencies measured by the test.

The target composite variable can be taken as the weighted sum of the two variables of interest, μ_1 and μ_2 and can be represented as:

$$\mu = \alpha_1\mu_1 + \alpha_2\mu_2$$

where α_1 and α_2 are the intended weights for the two types of competence being assessed. For convenience, we will assume that α_1 and α_2 sum to 1.0.

The target composite will be taken as the criterion in evaluating the validity of various observed composite scores. The value of this target composite is not known *a priori* for individual candidates and therefore cannot be used as an empirical criterion, but using the model assumptions and available estimates of observed score variances, covariance, and reliabilities, it is possible to estimate reliabilities and validities for different observed composites as measures of the target composite.

Validity of Observed Composite as an Estimate of the Target Composite

In estimating the target composite variable, μ, an observed composite,

$$X = w_1X_1 + w_2X_2$$

is employed, where w_1 and w_2 are not necessarily equal to α_1 and α_2.

The variance of the target composite is given by:

$$\sigma^2(\mu) = \alpha_1^2\sigma^2(\mu_1) + \alpha_2^2\sigma^2(\mu_2) + 2\alpha_1\alpha_2 cov_{12}$$

and the variance of the observed composite is given by:

$$\sigma^2(X) = w_1^2\sigma^2(X_1) + w_2^2\sigma^2(X_2) + 2w_1w_2 cov_{12}$$

The covariance between the observed composite and the target composite is given by:

$$cov(\mu,X) = \alpha_1 w_1 \sigma^2(\mu_1) + \alpha_2 w_2 \sigma^2(\mu_2) + (\alpha_1 w_2 + \alpha_2 w_1) cov_{12}$$

The validity coefficient for the observed composite, X, as a measure of the target composite is then given by:

$$\rho(X,\mu) = \frac{cov(\mu,X)}{\sigma(\mu)\sigma(X)}$$

or

$$\rho(X,\mu) = \frac{\alpha_1 w_1 \sigma^2(\mu_1) + \alpha_2 w_2 \sigma^2(\mu_2) + (\alpha_1 w_2 + \alpha_2 w_1) cov_{12}}{[w_1^2\sigma^2(X_1) + w_2^2\sigma^2(X_2) + 2w_1 w_2 cov_{12}]^{1/2}\,[\alpha_1^2\sigma^2(\mu_1) + \alpha_2^2\sigma^2(\mu_2) + 2\alpha_1\alpha_2 cov_{12}]^{1/2}}$$

The validity coefficient, $\rho(X,\mu)$, represents the accuracy with which the observed composite, X, measures the target composite, μ.

Reliability of the Observed Composite, X

The reliability of the observed composite is given by the ratio of the covariance between the observed scores obtained from two replications of the measurement procedure, over the product of their standard deviations. The covariance is given by:

$$cov(X,X') = w_1^2\sigma^2(\mu_1) + w_2^2\sigma^2(\mu_2) + 2w_1 w_2 cov_{12}$$

Note that this expression is the same as the variance of the target composite variable, except for the fact that α_1 is replaced by w_1 and α_2 is replaced by w_2. Since the scores from the two replications have the same expected variance, the reliability of the composite is given by:

$$\rho(X,X') = \frac{cov(X,X')}{\sigma^2(X)}$$

or

$$\rho(X,X') = \frac{w_1^2\sigma^2(\mu_1) + w_2^2\sigma^2(\mu_2) + 2w_1 w_2 cov_{12}}{w_1^2\sigma^2(X_1) + w_2^2\sigma^2(X_2) + 2w_1 w_2 cov_{12}}$$

The reliability of the observed composite can also be represented as the ratio of the true-score variance of the composite, X, to the observed-score variance of the composite. The true-score variance of the observed composite is not generally equal to the variance in the target composite, unless $w_1 = \alpha_1$ and $w_2 = \alpha_2$, and we will be considering cases where this is not the case. We will let μ_X represent the universe score corresponding to the observed composite score, to distinguish it from the target composite score, μ. The reliability of the observed composite score can then be represented as:

$$\rho(X,X') = \frac{\sigma^2(\mu_X)}{\sigma^2(X)} = \rho^2(X,\mu_X)$$

The composite universe score is equal to the covariance between two instances of the observed composite variable, X and X'.

Disattenuated Validity

Finally, a disattenuated correlation between the observed composite and the target composite can be represented as:

$$\rho(\mu_X,\mu) = \frac{\rho(X,\mu)}{\rho(X,\mu_X)}$$

The disattenuated validity coefficient provides an indication of what the correlation between the observed composite and the target composite would be if the observed composite could be measured without error. It corrects for the lack of perfect reliability in the observed composite.

Therefore, the validity coefficient can be represented as the product of the reliability of the observed composite score and the disattenuated validity coefficient:

$$\rho(X,\mu) = \rho(X,\mu_X)\,\rho(\mu_X,\mu)$$

That is, the inference from X to μ can be partitioned into two steps, an inference from X to μ_X and an inference from μ_X to μ.

Some Simulated Examples

Although the models used in this paper are basically quite simple (employing only the assumptions of classical test theory, and some assumptions about the intent of the test user), the resulting equations are fairly complicated. The implications of these equations for the choice of weights is not transparent. Therefore, in this section, we will use graphs of the equations for some values of the reliabilities of the two subtest scores, of their variances, and of their correlation, as well as two choices for the target score weights.

In particular, we will assume that we have the commonly occuring situation, in which the scores from a multiple choice (MC) test and an essay test are combined to produce a composite score. We will assume that the MC scores have a reliability of 0.90, and that the essay scores have a reliability of 0.72. The reliabilities of two tests place limits on the correlation between the two tests, and the reliability of 0.72 for the essay scores was chosen to be realistic, and yet to allow for an MC-essay correlation as high as 0.80.

The Reliability of the Observed Composite

The reliability of the observed composite score is represented in Figure 1 for three values of the correlation between the multiple-choice test scores and essay scores, 0.4, 0.6, and 0.8. Each of the three curves in Figure 1 represents the reliability of the observed composite scores for the full range of possible MC weights, from 0.0 to 1.0. The weights for the MC and essay test have a sum of 1.0; so, as the MC weight increases from 0.0 to 1.0, the weight for the essay scores decreases from 1.0 to 0.0. As noted above, in estimating the composite reliability as a function of the MC weight for Figure 1, the reliability of the MC was assumed to be 0.90, and the reliability of the essay test was assumed to be 0.72.

Note that the reliability of the observed composite does not depend on the values of the weights for the target composite. It depends only on the reliabilities of the two tests, their correlation, and the weights assigned to the two tests in the observed composite.

The three curves (corresponding to assumed MC-essay correlations of 0.4, 0.6, and 0.8) are close together, indicating that the reliability of the composite is not very sensitive to the correlation between the MC and essay tests. The reliability of the composite consistently increases as the correlation between the MC and essay scores increases, but the change is not very large.

The reliability of the composite is more sensitive to the weights assigned to the MC and essay scores. On the extreme left of Figure 1, where the MC weight is 0.0,and the essay weight is 1.0, the reliability is equal to 0.72, independent of the correlation between the MC and the essay scores. A weight of 0.0 for the MC implies that we are ignoring the MC scores and focusing exclusively on the essay scores and as a result, the reliability of the composite is simply the reliability of the essay scores. At the extreme right, the MC has a weight of 1.0 and the essays have a weight of 0.0, and the reliability that is equal that of the MC, 0.90. Between these two extremes, the reliability tends to increase as the weight assigned to the more reliable of the two measures, the MC test, increases.

Note, however, that the increase is not uniform. The composite reliability increases approximaely linearly as the MC weight increase from 0.0 to about 0.5. It then levels off and eventually decrease slightly as the weight assigned to the MC increases from 0.5 to 1.0. Assuming that the correlation between the MC scores and the essay scores is 0.6 (the value for the middle curve in Figure 1), the composite reliability is at or above 0.90 for any composite with an MC weight of 0.6 or above. The maximum reliability of about 0.91, which is achieved for an MC weight of about 0.8 and a corresponding essay weight of about 0.2, is not much higher than this value.

For MC and essay weights of 0.5 and a correlation of 0.6 between the MC and essay scores, the reliability of the composite is 0.88, which is still close to the maximum reliability of 0.91. For an MC weight of 0.4 and a correlation of 0.6 between the MC and essay scores, the reliability of the composite drops to about 0.85, and for still lower MC weights, the reliability drops substantially. These analyses of the reliability of the composite suggest that for tests with the properties assumed here, the weight assigned to the MC should be about 0.4 or higher, in order to maintain adequate reliability.

So Figure 1 suggests that if the goal is to maximize the reliability, it would make sense to assign very heavy weights to the MC score and relatively low weights to the essay scores, independent of the weights assigned to the MC and essay scores in the target composite. However, Figure 1 also indicates that, for tests with reliabilities of about 0.90 and 0.72 and a covariance of about 0.6, we can give the essay scores considerable weight without doing much damage to the reliability. As we shall see in the next section, validity can be enhanced by giving substantial weight to the essay scores.

Note that these analyses ignore an important potential option. If the decisions were made to assign the MC score a weight of 1.0 (and therefore to assign the essay score a weight of 0.0); there would be no reason to administer the essay test. In this context, the reliability could be strengthened further by increasing the length of the MC test. If the MC test length were doubled, the Spearman-Brown formula would predict that the reliability of the double-length MN test would be 0.947.

Validity of Observed Composite as an Estimate of the Target Composite

In estimating the target composite variable, an observed composite,

$$X_C = w_M X_M + w_E X_E$$

is taken as an estimate of the target composite,

$$T_C = \alpha_M T_M + \alpha_E T_E$$

where w_M and w_E are not necessarily equal to α_M and α_E.

The validity coefficient for the observed composite, X, as a measure of the target composite represents the accuracy with which the observed composite, X, measures the target composite.

The validity of the observed composite as a measure of the target true-score composite does depend on the values of the weights for the target composite as well as the weights for the observed composite. For illustration, we will consider two cases. Results for the case in which the MC scores and essay scores have the same weight in the target composite are presented in Figure 2a ($\alpha_M = \alpha_E = 0.5$). Results for the case in which the essay scores are weighted twice as heavily as the MC scores in the target composite are presented in Figure 2b ($\alpha_M = 0.33$ and $\alpha_E = 0.67$). Policy makers would define the relative importance of the MC and essay tests by specifying the weights in the target composite.

Figures 2a and 2b present the validity coefficients as a function of the weight assigned to the MC in the observed composite scores over the full range of possible MC weights, from 0.0 to 1.0. Using a format analogous to Figure 1, Figures 2a and 2b present curves corresponding to three values of the correlation between MC scores and essay scores, 0.4, 0.6, and 0.8. Each of the three curves in Figures 2a and 2b represent the validity coefficient for the observed composite scores as a function of MC weights ranging from 0.0 to 1.0. As noted earlier, as the MC observed-score weight increases from 0.0 to 1.0, the weight for the essay scores decreases from 1.0 to 0.0.

In estimating the validity coefficients in Figures 2a and 2b, the reliability of the MC was again assumed to be 0.90, and the reliability of the essay test was assumed to be 0.72. The validity coefficients as a function of MC weight are reported separately for different values of the correlation between the MC and essay scores. The middle curve reports the validity as a function of MC weight assuming that the correlation between MC and essay scores is 0.6, which is a typical value for this correlation. The upper curve reports the validity as a function of MC weight assuming that the correlation between MC and essay scores is 0.8, which is a very high value for this correlation (about as high as it can be, given the reliabilities of the MC and essay tests). The lower curve assumes a relatively low MC-essay correlation of 0.4.

For the sake of brevity, we will focus our discussion of Figures 2a and 2b on the middle curve, which assumes a value of 0.6 for the correlation between MC scores and essay scores. The other two curves provide an indication of how this relationship between the validity and the MC weight is likely to vary as the MC-essay correlation varies. Note that for both Figure 2a and Figure 2b, the validity coefficients consistently increase as the correlation between the MC and essay scores increases from 0.4 to 0.8., and the change can be quite large.

In Figure 2a, the weight of the MC scores was assigned to be equal to the weight of the essay scores in the target true-score composite. This weighting would make sense if the competencies measured on the MC and the essay test were equally valued by the policy makers. The curve of validity as a function of the MC weight in the observed composite scores, assuming a value of 0.6 for the MB-essay correlation, increases to a maximum of about 0.95, and then decreases for higher values of w_M. This curve hits its maximum when the MC weight is about 0.63 and the essay weight is about 0.37. That is, even though the essay and MC have equal weights in the target true-score composite, the optimal observed score composite assigns the MC scores a higher weight than the essay scores. However, the optimal weight for the MC using validity as the benchmark is considerably lower than it would be if reliability were the sole concern. Furthermore, the validity coefficient has a value of at least 0.90 over a wide range of MC weights ranging from a low of about 0.34 to a high of about 0.94. So, assuming that a weighting choice is acceptable as long as it generates a composite validity of 0.90 or above, policy makers have considerable flexibility in weighting the two scores in the observed composite.

There are three variables that influence the relationship between the validity coefficient and the weights assigned to the MC and essay scores. First, all else being equal, the validity of the observed composite score as a measure of the target true-score composite will tend to be higher to the extent that the weights in the observed composite match the weights in the target composite score. In Figure 2a, it is assumed that, in the target composite score, the weights of the MC and essay scores are equal. To get a good estimate of this composite, it is necessary to include measures of the competencies assessed by the MC and the competencies measured by the essay test. If the reliabilities of the MC and essay tests were the same, the optimal weighting for the observed score composite would be the same as that of the target composite.

The second major influence on the validity coefficients is the relative values of the reliabilities of the two tests. To the extent that one of the reliabilities (i.e., the MC) is substantially higher than that of the other test (i.e., the essay test), the reliability of the composite tends to increase as more weight is given to the more reliable test (see Figure 1). All else being equal, as the reliability of the composite increases, the validity will increase. The net effect of the difference in reliability between the MC and essay tests in Figure 2a is to push the maximum validity away from the center of the graph, where the weights are equal, toward the right side of the graph, where the MC gets more weight.

The third influence on the validity coefficients is the correlation between MC scores and essay scores. As noted earlier, as the MC-essay correlation goes up, the validity coefficients consistently increase. This is due in part to the fact that as the MC-essay correlation increases, the reliability of the observed composite increases. In addition, as this correlation increases, the MC and essay scores tend to become interchangeable from a statistical point of view. That is, as the MC-essay correlation increases, the MC scores get better at estimating essay scores, and the essay scores get better at estimating MC scores. As a result, it becomes less important that the weights in the observed composite match those of the target composite, and to the extent that this happens, it is statistically advantageous to lean more heavily on the more reliable of the two tests. This is typically the MC test. So, for example, in Figure 2a, as the MC-essay correlation increases from 0.4 to 0.6 to 0.8, the optimal weighting of the MC increases from approximately 0.58 to 0.67 to 0.75.

Figure 2b is analogous to Figure 2a except that the essay test is assumed to get twice as much weight as the MC in the target true-score composite. This weighting would make sense if the competencies measured on the essay test were considered to be about twice as important as those measured by the MC.

The curve of validity as a function of the MC weight in the observed composite scores, assuming a value of 0.6 for the MC-essay correlation, hits a maximum of about 0.93 when the MC weight is about 0.53 and the essay weight is about 0.47. However, the validity coefficient has a value of at least 0.90 over a wide range of MC weights from about 0.30 to about 0.76.

In Figure 2b, it is assumed that the target composite score weights the true essay scores twice as heavily as the true MC scores. If the reliabilities of the MC and essay tests were the same, the optimal weighting for the observed score composite would also weight the essay scores twice as heavily as the MC scores. However, the difference in reliability between the MC and essay tests tends to push the maximum validity toward the right, where the MC gets more weight. In this case the net result is that the MC and essay scores get approximately equal weights in the optimal observed-score composite.

As the MC-essay correlation increases, it is statistically advantageous to give more emphasis to the MC. In Figure 2b, as the MC-essay correlation increases from 0.4 to 0.6 to 0.8, the optimal weighting of the MC increases from approximately 0.45 to 0.53 to 0.75.

The point of these graphs is not to identify an optimal weighting of the two test scores in the composite, but to indicate the general trends in the data. In using these results, it is important to keep the surrogation fallacy in mind. The fact that one test can substitute for another test in estimating a target score does not imply that the two tests have the same meaning or that the two test are interchangeable in terms of their consequences. To the extent that the test content tends to drive instruction and learning, it may be desirable to include essays and other types of performance tests within high-stakes testing programs. The results in this section suggest that, at least for the cases considered here, policy makers have considerable latitude in assigning the weights, while maintaining adequate reliability and validity.

Errors of Measurement

The total error variance in estimating the target composite from the observed composite can be estimated from the validity coefficient, which is given by the correlation between X and μ, and the estimated variance in μ. In particular, the proportion of the variance in X accounted for by μ is given by the coefficient of determination, $\rho^2(X, \mu)$. Therefore the error variance in using X to estimate μ is

$$\begin{aligned}\sigma^2(e_{tot}) &= \sigma^2(X)[1 - \rho^2(X,\mu)] \\ &= \sigma^2(X)[1 - \sigma^2(\mu)/\sigma^2(X)] \\ &= \sigma^2(X) - \sigma^2(\mu)\end{aligned}$$

This total error can be partitioned into two parts. The first part is the random error involved in generalizing from the observed score, X, to the universe scope, μ_x, the expected value of the observed composite score over replications of the measurement procedure using the weights, w_1, and w_2. The random error associated with the composite variable, X, represents the error in estimates of μ_X from X. The second part of the error is the systematic error involved in inferences from μ_X to μ, the target composite.

The reliability coefficient, $\rho^2(X, \mu_x)$, provides an estimate of the percentage of observed score variance attributable to true score variance. Therefore, the variance of the random error, e_{ran}, is given by:

$$\begin{aligned}\sigma^2(e_{ran}) &= \sigma^2(X)[1 - \rho^2(X,\mu_X)] \\ &= \sigma^2(X)[1 - \sigma^2(\mu_X)/\sigma^2(X)] \\ &= \sigma^2(X) - \sigma^2(\mu_X)\end{aligned}$$

This error variance represents the random fluctuations in the composite, X, due to the random fluctuations in the two observed variables, X_1 and X_2. The weights, w_1 and w_2, are considered fixed.

The squared disattenuated validity coefficient, $\rho^2(\mu, \mu_x)$ provide an estimate of the proportion of variance in μ_x attributable to the target composite variable, μ. Therefore the systematic errors, e_{sys}, rising from the use of the weights, w_1, and w_2, instead of the intended weights, α_1 and α_2, is given by:

$$\begin{aligned}\sigma^2(e_{sys}) &= \sigma^2(\mu_X)[1 - \rho^2(\mu_X,\mu)] \\ &= \sigma^2(\mu_X)[1 - \sigma^2(\mu)/\sigma^2(\mu_X)] \\ &= \sigma^2(\mu_X) - \sigma^2(\mu)\end{aligned}$$

Given these results, it is clear that, as expected:

$$\sigma^2(e_{tot}) = \sigma^2(e_{sys}) + \sigma^2(e_{ran})$$

Therefore, weights that maximize reliability, and therefore minimize $\sigma^2(e_{ran})$, do not necessarily minimize the total error variance or maximize the accuracy of estimates of the target composite. In order to effectively control the total error, it is necessary to control both the systematic errors due to the use of weights, w_1, and w_2, which are different from the target weights, and random errors associated with a lack perfect reliability. As indicated in this paper, these two goals can be at odds in some cases.

Focusing exclusively on reliability tends to generate good estimates of the universe score, μ_X, which may be quite different from the actual target variable. If μ_1 and μ_2 are highly correlated, μ_X will be highly correlated with μ, the observed score composite that optimizes the reliability may also provide excellent validity. If the correlation between μ_1 and μ_2 is not close to 1.0, the composite that optimizes reliability may also introduce a lot of systematic error. Reducing random error by increasing systematic error is generally not a good tradeoff unless the reduction in the magnitude of random errors is large compared to the increases in the magnitude of the systematic error.

Results

Policy makers would define the relative importance of the component measures by specifying the weights in the target composite. In this paper, we consider two possible target composites, one in which the two measures have equal weights, and one in which the less reliable measure has twice the weight of the more reliable measure. Several sets of analyses with different values for the component reliabilities, variances, covariances, and target weights are reported. These analyses illustrate several general conclusions (assuming that one component test has a substantially higher reliability than the other, e.g., 0.9 vs. 0.72, that the two tests have similar variances and a modest to fairly high correlation, e.g., 0.4 to 0.8.

1. The reliability of the observed composite does not depend on the definition of the target composite. It depends only on the reliabilities of the two component tests, their variances and correlation, and the weights assigned to the two tests in the observed composite. The reliability of the composite tends to increase as the correlation between the two component scores increases. The reliability of the composite also tends to increase as the weight assigned to the more reliable of the two measures increases, but levels off and decreases slightly as the weight assigned to the more reliable measure approaches 1.0. Consistent with Wainer and Thissen (1993), if the goal is to maximize the reliability, it makes sense to assign almost all of the weight to the more reliable of the two observed scores, even if the two components have equal weights in the target composite. However, the analyses also indicate that we can give the essay scores considerable weight without doing much damage to the reliability. In addition, of course, if little or no weight is to be given to the less reliable of the two measures, substantial savings in time and money could be realized; or the reliability could be further increased by lengthening the more reliable test.

2. The validity of the observed composite as a measure of the target true-score composite does depend on the match between the weights for the target composite and the weights for the observed composite. The validity of the observed composite score as a measure of the target true-score composite will tend to be higher to the extent that the weights in the observed composite match the weights in the target composite score. If the reliabilities of the component scores were the same, the optimal weighting for the observed score composite would be the same as that of the target composite.

3. To the extent that reliability of one of the tests is substantially higher than that of the other test, the reliability of the composite tends to increase as more weight is given to the more reliable test. This increases the reliability of the observed composite, and all else being equal, as the reliability of the composite increases, the validity will increase. As a result, to the extent that one of the observed scores is more reliable than the other, validity is enhanced by giving somewhat more weight to the more reliable test than is indicated by the target weights.

4. As the correlation between the component tests increases, the validity coefficients increase, due in part to the fact that the reliability of the observed composite increases. In addition, as the more reliable score gets better at estimating the less reliable score it is less important that the weights in the observed composite approximate those of the target composite, and it is advantageous to lean more heavily on the more reliable of the two tests. Even if the less reliable test is assigned twice the weight of the more reliable test in the target true-score

composite, and the correlation between the two tests is fairly high (e.g., around 0.6), the maximum validity occurs when the weights in the observed composite are approximately equal.

These results provide general guidance in defining the weights to be assigned to two tests (and can be easily extended to three or more component tests). They suggest that, if the target variable includes two distinct components, it will generally be important to measure both components in order to achieve high validity, even if one of the measures is less reliable than the other. Giving extra weight to the more reliable observed score will improve the reliability of the composite, and up to a point will tend to improve its validity. Giving too much weight to the more reliable component decreases the validity of the observed composite as a measure of the target composite.

References

Bennett, R., Rock, D., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. journal of educational measurement, 28, 77-92.

Brennan, R. (2001). Some problems, pitfalls, and paradoxes in educational measurement. Educational measurement: Issues and Practice, 20, 4, 6-18

Feldt, L. (1997). Can validity rise when reliability declines? Applied Measurement in Education 10, 377-387.

Gulliksen, H. (1950). Theory of Mental tests. New York, John Wiley.

Hambleton, R. & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. Applied Measurement in Education, 10, 19-38.

Jarjoura, D. & Brennan, R. (1982). A variance components model for measurement procedures associated with tables of specifications. Applied Psychological Measurement, 6,2,161-171.

Kelly, T. (1927). Interpretation of Educational measurements. New York, World book Company.

Klein, S. (1995). Options for combining MBE and essay scores. The Bar Examiner, 38-43.

Lenel, J. (1992) Issues in equating and combining MBE and essay scores. The Bar Examiner, 6-20

Mehrens, W. (1990) Combining evaluation data from multiple sources. In J. Millman & L. Darling-Hammond (Eds.), The New hanbook of teacher Evaluation. Sage, Newbury park, CA.

Rudner, L. Informed test component weighting. Educational measurement; Issues and Practice, 20, 1, 16-19.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. Applied Measurement in Education, 6, 103-118.

Wang M. & Stanley J. (1970). Differential weighting: A review of methods and empirical studies. Review of Educational Research, 40, 663-705.

Fig. 1: Reliability vs. MC wt.

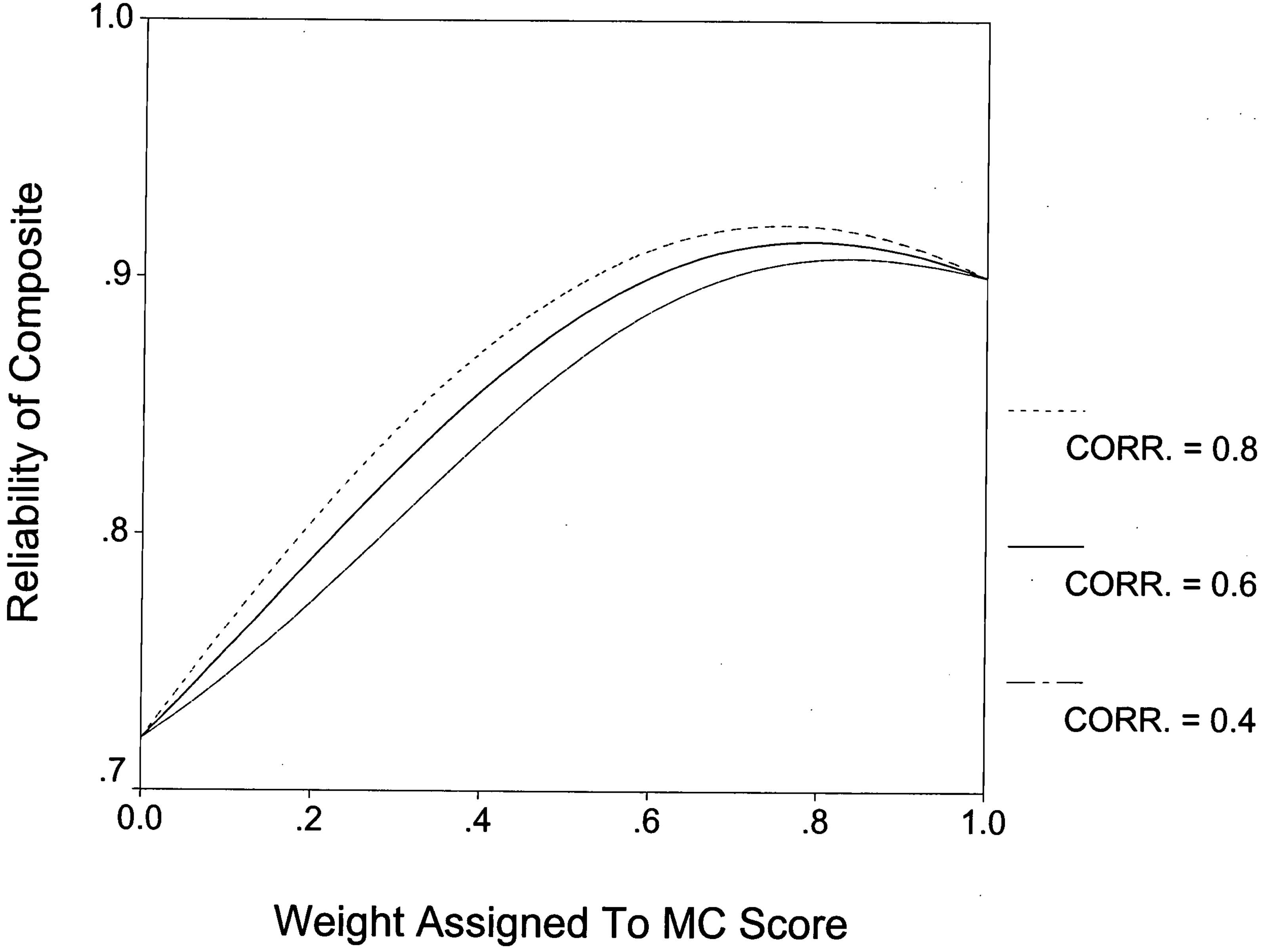


Fig. 2a: Validity vs. MC Test wt.

Target Essay wt. = Target MC wt.

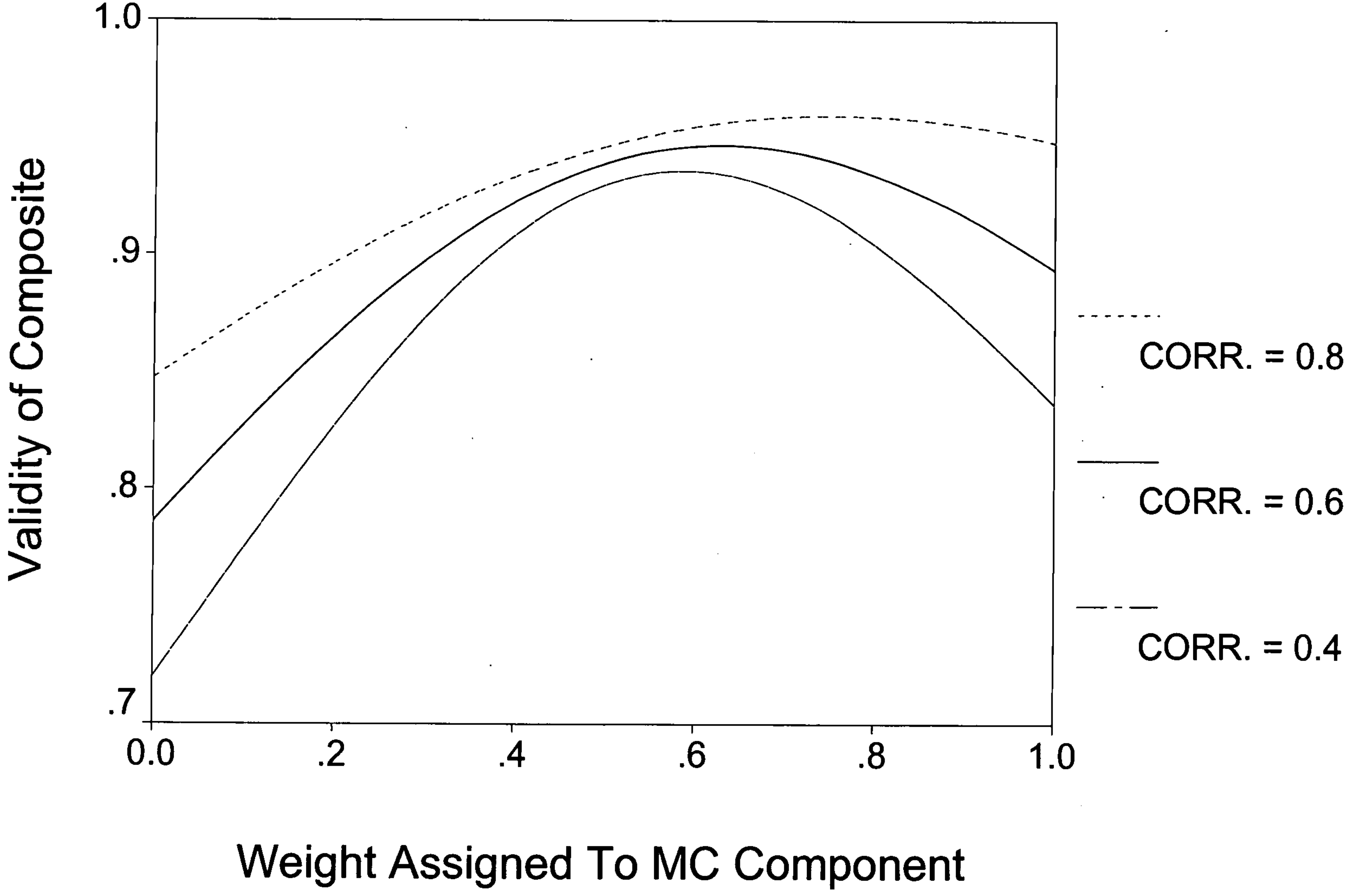
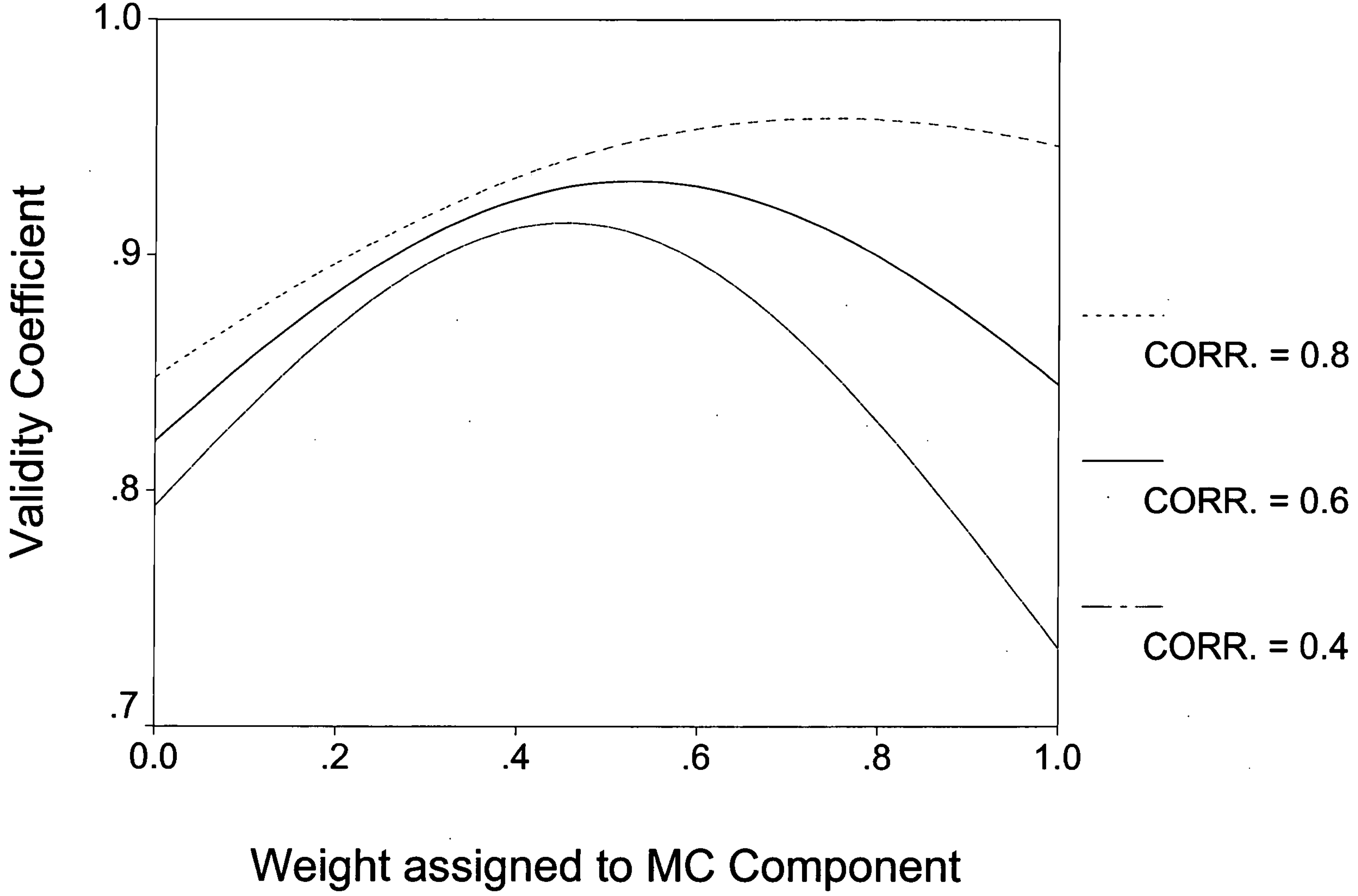


Fig. 2b: Validity vs. MC Test wt.

Target Essay wt. = 2 xTarget MC wt.



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034923

I. DOCUMENT IDENTIFICATION:

Title: The Reliability and Validity of Weighted Composite Scores	
Author(s): Michael Kane, Susan Case	
Corporate Source: National Conference of Bar Examiners	Publication Date: March 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be effixed to all Level 1 documents	The sample sticker shown below will be effixed to ell Level 2A documents	The sample sticker shown below will be effixed to ell Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☑	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 releese, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Michael T. Kane	Printed Name/Position/Title: MICHAEL T. KANE	
Organization/Address: National Conference of Bar Examiners	Telephone: 608-280-8550	FAX:
	E-Mail Address: mkane@ncbex.org	Date: 5/12/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse: **University of Maryland**
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Lab, Bldg 075
College Park, MD 20742
Attn: Acquisitions