Agreement Coefficients as Indices of Dependability for Domain-Referenced Tests

Michael T. Kane
National League for Nursing

Robert L. Brennan
American College Testing Program

A large number of seemingly diverse coefficients have been proposed as indices of dependability, or reliability, for domain-referenced and/or mastery tests. In this paper it is shown that most of these indices are special cases of two generalized indices of agreement—one that is corrected for chance and one that is not. The special cases of these two indices are determined by assumptions about the nature of the agreement function or, equivalently, the nature of the loss function for the testing procedure. For example, indices discussed by Huynh (1976), Subkoviak (1976), and Swaminathan, Hambleton, and Algina (1974) employ a threshold agreement, or loss, function; whereas, indices discussed by Brennan and Kane (1977a, 1977b) and Livingston (1972a) employ a squared-error loss function. Since all of these indices are discussed within a single general framework, the differences among them in their assumptions, properties, and uses can be exhibited clearly. For purposes of comparison, norm-referenced generalizability coefficients are also developed and discussed within this general framework.

Glaser and Nitko (1971) define a criterion-referenced test as "one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 653). Hively (1974) and Millman (1974), among others, suggest the term "domain-referenced test" rather than "criterion-referenced test." Noting that the word "criterion" is ambiguous in some contexts, they have argued that the word "domain" provides a more direct specification of the entire set of items or tasks under consideration. A *mastery test* can be defined as a domain-referenced test with a single cutting score.

One can also distinguish between a particular type of test (norm-referenced or domain-referenced) and the scores (or interpretation of scores) resulting from a test. For example, the scores from any test might be given norm-referenced or domain-referenced interpretations. Indeed, most of the literature on the dependability (or reliability) of domain-referenced tests actually treats the dependability of a particular set of scores that are given a domain-referenced or mastery interpretation. In this paper, to obviate verbal complexity, there will often be references to norm-referenced, domain-referenced, and mastery "tests"; however, a more complete verbal description would refer to scores that are given norm-referenced, domain-referenced, or mastery interpretations for a particular testing procedure.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 1 Winter 1980 pp. 105–126
© Copyright 1980 West Publishing Co.

Since Popham and Husek (1969) challenged the appropriateness of correlation coefficients as indices of reliability for domain-referenced and mastery tests, considerable effort has been devoted to developing more appropriate indices. Most of these indices have been proposed as measures of *reliability*; however, in this paper the more generic term, *dependability,* will be used in order to avoid unwarranted associations with the classical theory of reliability for norm-referenced tests.

A large number of seemingly diverse coefficients have been proposed (see Hambleton, Swaminathan, Algina, & Coulson, 1978). In this paper, it is shown that most of these indices can be classified into two broad categories, depending on their underlying (and sometimes unstated) assumptions about the nature of agreement or, equivalently, the nature of loss in the testing procedure. For example, indices discussed by Carver (1970), Huynh (1976), Subkoviak (1976), and Swaminathan, Hambleton, and Algina (1974) employ a threshold agreement, or loss, function; whereas, indices discussed by Brennan and Kane (1977a, 1977b) and Livingston (1972a, 1972b, 1972c, 1973) employ a squared-error loss function. Recently, Subkoviak (in press) has reviewed some threshold agreement indices and Brennan (in press) has reviewed some indices involving squared-error loss.

A Pair of General Indices of Dependability

Agreement Function

A general expression for the dependability of a testing procedure can be derived by examining the expected agreement between two randomly selected instances of a testing procedure. Any particular instance of a testing procedure will be referred to as a "test." No assumptions need to be made about the nature of the tests, the details of their administration, or their scoring. Since the instances, or tests, are randomly selected from a *universe* of possible instances, they are randomly parallel. Therefore, the *expected* distribution of outcomes for the population is assumed to be the same for each administration of the testing procedure. This does not imply that the distributions of outcomes are necessarily identical for all tests; that is, the stronger assumption of classically parallel tests is not made.

The degree of agreement between any two scores, s_i and s_j, is defined by an agreement function, $a(s_i, s_j)$. The scores, s_i and s_j, may be raw scores or they may be transformed in some way. For convenience, it will be assumed that only a finite number of scores ($s_o, \ldots, s_i, \ldots, s_n$) may result from the use of the testing procedure. The form of the agreement function defines what is meant by agreement in any particular context. In general, the agreement function will reflect intuitive notions of the relative degree of agreement for different pairs of scores.

Although no particular form will be assumed for an agreement function in the development of general indices of dependability, it is reasonable to impose some conditions on the class of functions that will be accepted as agreement functions. It will be assumed that all agreement functions satisfy the following three conditions:

(i) $a(s_i, s_i) \geq 0;$ [1a]

(ii) $a(s_i, s_j) = a(s_j, s_i);$ and [1b]

(iii) $a(s_i, s_i) + a(s_j, s_j) \geq 2a(s_i, s_j).$ [1c]

Given that the agreement between randomly parallel tests is being examined, the first two of these conditions are certainly natural. The third condition implies that the agreement assigned to any pair of scores, s_i and s_j, cannot be greater than the average of the agreements that results from pairing each of these scores with itself. All the agreement functions discussed in this paper satisfy these three conditions.

Maximum Agreement and the Index θ

The score for person v on the k^{th} instance of the testing procedure can be represented by the random variable S_{vk}. Similarly, S_{wl} is the score for person w on test l. For every person v and every test k, S_{vk} takes one of the values $s_o, \ldots, s_i, \ldots, s_n$. The expected agreement given by

$$A = \underset{v,k,l}{\mathcal{E}} \; a(S_{vk}, S_{vl}), \qquad [2]$$

might then be considered as an index of dependability, where expectation is taken over the population of persons and over pairs of tests that are independently sampled from the universe of tests. The expected agreement may also be represented in terms of the joint distribution of scores on the two tests:

$$A = \sum_{i,j=0}^{n} a(s_i, s_j) \cdot Pr(S_{vk} = s_i, S_{vl} = s_j) \qquad [3]$$

where $Pr(S_{vk} = s_i, S_{vl} = s_j)$ is the probability that a randomly chosen person, v, obtained scores s_i and s_j on randomly chosen tests, k and l. Equations 2 and 3 represent the same quantity expressed in two different ways. In the following discussion, whichever of these expressions is most convenient for the issue under consideration will be used. The notation in Equation 3 can be simplified by letting

$$a_{ij} = a(s_i, s_j) \qquad [4]$$

and

$$P_{ij} = Pr(S_{vk} = s_i, S_{vl} = s_j). \qquad [5]$$

Equation 3 can then be written as

$$A = \sum_{i,j=0}^{n} a_{ij}p_{ij}. \qquad [6]$$

However, the index A depends on the scale chosen for a_{ij} and can be made arbitrarily large by multiplying a_{ij} by a sufficiently large constant. One way to correct this problem is to take as the index of agreement

$$\theta = \frac{A}{A_m}. \qquad [7]$$

In Equation 7, A_m is the expected agreement between an instance of the testing procedure and itself:

$$A_m = \underset{v,k}{\mathcal{E}} \; a(S_{vk}, S_{vk}) \qquad [8a]$$

$$= \sum_{i=0}^{n} a(s_i, s_i) \cdot \Pr(S_{vk} = s_i); \qquad [8b]$$

and in the simpler notation,

$$A_m = \sum_{i=0}^{n} a_{ii}p_i \;, \qquad [9]$$

where p_i is the probability that a randomly selected person will get the score s_i on a randomly chosen instance of the testing procedure. A is equal to A_m when each person in the population has the same score on every test, i.e., when all instances of the testing procedure are in perfect agreement in the assignment of scores, s_i, to persons in the population.

Using the three conditions in Equation 1, it is easy to show that for any marginal distribution, A_m is the maximum value of A. Since p_i is a marginal probability,

$$A_m = \sum_{i} a_{ii}p_i = \sum_{i,j} a_{ii}p_{ij} \;, \qquad [10]$$

and

$$A_m = \sum_{j} a_{jj}p_j = \sum_{i,j} a_{jj}p_{ij} \;. \qquad [11]$$

Therefore, A_m can be written as

$$A_m = \sum_{i,j} \left[\frac{a_{ii} + a_{jj}}{2} \right] p_{ij} \;. \qquad [12]$$

Now, using Assumption iii given by Equation 1c,

$$A_m \geq \sum_{i,j} a_{ij}p_{ij} \;; \qquad [13]$$

and using Equation 6

$$A_m \geq A \;. \qquad [14]$$

From the definition of θ in Equation 7, therefore, it follows that θ is less than or equal to one.

Chance Agreement and the Index θ_c

The index θ does not consider the contribution of chance agreement to the dependability of measurement. As will be seen, θ may be large even when scores are randomly assigned to persons on each test. When a score is assigned *by chance* to examinee v, this means that the score is randomly selected from the distribution of scores for the *population* of persons on the *universe* of tests. The assignment of s_i to examinee v by chance depends only on the marginal probability, p_i, of the score s_i and not on the examinee's performance. Therefore, for chance assignment the score assigned to an examinee on any particular test is independent of the score assigned on any other test.

The effect of chance agreement can be examined by taking the expected agreement between the score S_{vk} for person v on the k^{th} test and the score S_{wl} for an independently sampled person, w, on an independently sampled test, l:

$$A_c = \underset{v,w,k,l}{\mathcal{E}} \; a(S_{vk}, S_{wl}) \qquad [15]$$

or

$$A_c = \sum_{i,j=0}^{n} a(s_i, s_j) \cdot Pr(S_{vk} = s_i, S_{wl} = s_j) \; . \qquad [16]$$

Since both persons v and w and tests k and l are sampled independently,

$$Pr(S_{vk} = s_i, S_{wl} = s_j) = Pr(S_{vk} = s_i) \cdot Pr(S_{wl} = s_j) \; , \qquad [17]$$

where $Pr(S_{vk} = s_i)$ is the marginal probability that a randomly selected person will get the score s_i on a randomly chosen test. Substituting Equation 17 in Equation 16 and using the simplified notation introduced earlier gives

$$A_c = \sum_{i,j=0}^{n} a_{ij} p_i p_j \; . \qquad [18]$$

For any agreement function, Equation 18 depends only on the expected distribution of scores for a single administration of the testing procedure. A_c is the expected agreement for pairs of scores when each score is independently sampled from the marginal distribution of the population.

A general index of dependability, corrected for chance, can then be defined as

$$\theta_c = \frac{A - A_c}{A_m - A_c} \; . \qquad [19]$$

The numerator of Equation 19 provides a measure of how much the expected agreement for the testing procedure exceeds the expected agreement due to chance. Note that the joint distribution used to define A_c has the same marginal distribution as A; therefore, $A_m \geqslant A_c$. Also, $A_m \geqslant A$. Therefore,

$$A_m - A_c \geq A - A_c \; , \qquad [20]$$

the denominator of Equation 19 is the maximum value of the numerator, and θ_c is less than or equal to one.

Disagreement or Loss

Although most of the discussion in this paper is concerned with agreement functions, it will be useful in some places to discuss disagreement or loss. The expected loss, L, for any testing procedure can be defined as the difference between the maximum expected agreement and the expected agreement:

$$L = A_m - A \,. \qquad [21]$$

Using this definition, Equation 7 can be written:

$$\theta = \frac{A}{A + L} \,; \qquad [22]$$

and Equation 19 can be written:

$$\theta_c = \frac{A - A_c}{(A - A_c) + L} \,. \qquad [23]$$

Note that Equations 22 and 23 have the form of reliability or generalizability coefficients, with L taking the place of error variance (see Cronbach, Gleser, Nanda & Rajaratnam, 1972).

Interpretation of θ and θ_c

The two indices θ and θ_c address different questions about dependability. Although the properties of these indices will be discussed more fully later in the context of particular agreement functions, a brief statement is appropriate here. θ indicates how closely, in terms of the agreement function, the scores for an examinee can be expected to agree. θ_c indicates how closely (again, in terms of the agreement function) the scores for an examinee can be expected to agree, with the contribution of chance agreement removed. In other words, the index θ characterizes the dependability of decisions based on the testing procedure; whereas, the index θ_c characterizes the contribution of the testing procedure to the dependability of the decisions, over what would be expected on the basis of chance agreement. The two indices provide answers to different questions.

θ and θ_c for Threshold Agreement

Threshold Agreement Function

One common use of tests is to classify examinees into mutually exclusive categories. If there are only two categories, or if the categories are unordered, then a plausible agreement function for the classification procedure is the *threshold agreement function, t*:

$$t(S_{vk}, S_{wl}) = \begin{cases} 1 \text{ if } S_{vk} = S_{wl} \\ 0 \text{ if } S_{vk} \neq S_{wl} \end{cases}, \qquad [24]$$

where S_{vk} is the score (in this case the category) for examinee v on the test k. Equation 24 can be expressed more succinctly as

$$t_{ij} = t(s_i, s_j) = \begin{cases} 1 \text{ if } s_i = s_j \\ 0 \text{ if } s_i \neq s_j \end{cases}, \qquad [25]$$

where the score s_i represents assignment to the i^{th} category. Equation 25 has the advantage of notational simplicity; whereas, Equation 24 is a more detailed statement of the threshold agreement function, t. For either expression, the assigned agreement is one if the examinee is placed in the same category on both administrations of the procedure, and the agreement is zero if the examinee is placed in different categories on the two administrations. It is easily verified that the threshold agreement function satisfies the three conditions in Equation 1.

The Index θ(t)

Substituting t_{ij} for a_{ij} in Equation 6, the expected agreement is

$$A(t) = \sum_{i,j} t_{ij}p_{ij} = \sum_{i} p_{ii} . \qquad [26]$$

The maximum agreement is given by

$$A_m(t) = \sum_{i} t_{ii}p_i = 1 . \qquad [27]$$

Then, from Equation 7, the index of dependability, θ, for the threshold agreement function is

$$\theta(t) = \sum_{i} p_{ii} . \qquad [28]$$

Equation 28 states that the dependability of the classification procedure is simply the probability that a randomly chosen examinee will be placed in the same category on two randomly chosen tests. Note that $A(t)$, $A_m(t)$, and $\theta(t)$ are all equal to one if the classification procedure consistently places all examinees into a single category.

Equation 28 is stated in terms of population parameters. Estimates of $\theta(t)$, based on two administrations of the testing procedure, have been discussed by Carver (1970) and Swaminathan et al. (1974). Estimates of $\theta(t)$, based on a single administration of the testing procedure, have been discussed by Subkoviak (1978). Both types are reviewed by Subkoviak (in press).

The Index θ_c (t)

For the threshold agreement function in Equation 25, the expected agreement due to chance is

$$A_c(t) = \sum_{i,j} t_{ij} p_i p_j = \sum_i p_i^2 . \qquad [29]$$

From Equation 19, therefore, the index of dependability corrected for chance for a threshold agreement function is

$$\theta_c(t) = \frac{\Sigma p_{ii} - \Sigma p_i^2}{1 - \Sigma p_i^2} . \qquad [30]$$

In the special case where all examinees are consistently placed in a single category, $A_c(t)$ is equal to one and $\theta_c(t)$ is indeterminate.

The index $\theta_c(t)$ in Equation 30 is identical to Cohen's (1960) coefficient kappa and to Scott's (1955) coefficient, under the assumption that the expected score distributions for the two instances of the testing procedure are identical. As such, $\theta_c(t)$ has been proposed by Huynh (1976) and Swaminathan et al. (1974) as an index of reliability for mastery tests with a single cutting score.

Threshold Loss

The loss associated with a threshold agreement function can be determined by subtracting Equation 26 from Equation 27:

$$L(t) = A_m(t) - A(t) = \sum_{i \neq j} p_{ij} . \qquad [31]$$

If the two instances of the testing procedure assign the person to the same category, the loss is zero. If the two instances assign a person to different categories, the loss is one, regardless of which categories are involved. This is consistent with the usual definition of a threshold loss function (see Hambleton & Novick, 1973).

Interpretation of θ(t) and θ_c(t)

The first block of Table 1 summarizes results for the parameters $A(t)$, $A_m(t)$, $A_c(t)$, and $L(t)$ and the agreement indices $\theta(t)$ and $\theta_c(t)$ for the threshold agreement function, t.

As noted earlier, $\theta(t)$ will be equal to one whenever all instances of the procedure place everyone in the same category. The testing procedure used to assign persons to categories is then perfectly dependable. However, once it is established that all, or almost all, persons fall into one category, there is little to be gained by administering tests.

If almost everyone is in one category, the expected chance agreement, $A_c(t)$, will be close to $A_m(t)$, the maximum expected agreement. Under these circumstances, it would be difficult for any testing procedure to provide much improvement over chance assignment. Consequently, $\theta_c(t)$ will tend to be small whenever almost everyone is placed in the same category.

Therefore, $\theta_c(t)$ is open to the objection raised by Popham and Husek (1969) against the use of classical reliability coefficients with mastery tests—namely, $\theta_c(t)$ may be close to zero even when individuals are consistently placed in the correct category. However, this is not a flaw in the coefficient, but rather a possible source of misinterpretation. A low value of $\theta_c(t)$ does not necessarily indicate that assignments to categories are inconsistent from one administration to the next. Rather, a low value of

Table 1

Coefficients for Different Agreement Functions

Agreement Function	Parameters	Agreement Coefficients
Threshold	$A(t) = \Sigma p_{ii}$	$\theta(t) = \Sigma p_{ii}$
$t(S_{vk}, S_{wl}) =$	$A_m(t) = 1$	$\theta_c(t) = \dfrac{\Sigma p_{ii} - \Sigma p_i^2}{1 - \Sigma p_i^2}$
$\begin{cases} 1 \text{ if } S_{vk} = S_{wl} \\ 0 \text{ if } S_{vk} \neq S_{wl} \end{cases}$	$A_c(t) = \Sigma p_i^2$ $L(t) = \Sigma p_{ij} \quad (i \neq j)$	
Domain-Referenced	$A(d) = (\mu - \lambda)^2 + \sigma^2(\pi)$	$\theta(d) = \dfrac{(\mu - \lambda)^2 + \sigma^2(\pi)}{(\mu - \lambda)^2 + \sigma^2(\pi) + \sigma^2(\Delta)}$
$d(S_{vI}, S_{wJ}) =$	$A_m(d) = (\mu - \lambda)^2 + \sigma^2(\pi)$	
$(S_{vI} - \lambda)(S_{wJ} - \lambda)$	$+ \sigma^2(\Delta)$	$\theta_c(d) = \dfrac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\Delta)}$
	$A_c(d) = (\mu - \lambda)^2$	
	$L(d) = \sigma^2(\Delta)$	
Norm-Referenced	$A(g) = \sigma^2(\pi)$	$\theta(g) = \theta_c(g) = \dfrac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\delta)}$
$g(S_{vI}, S_{wJ}) =$	$A_m(g) = \sigma^2(\pi) + \sigma^2(\delta)$	
$(S_{vI} - \mu_I)(S_{wJ} - \mu_J)$	$A_c(g) = 0$	
	$L(g) = \sigma^2(\delta)$.	

$\theta_c(t)$ indicates that the testing procedure is not much more dependable in classifying individuals than a process of random assignment based on prior information about the population (i.e., the marginals in the population). Since $\theta(t)$ is large whenever the classification of examinees is consistent from one instance of the testing procedure to another, it is not subject to Popham and Husek's objection.

θ and θ_c for Domain-Referenced Agreement

In the following discussion of domain-referenced agreement, it will initially be assumed that, for each instance of the testing procedure, a random sample of *n* items is drawn from some infinite domain (or universe) of items and is administered to all examinees (i.e., items are crossed with persons).

The score for person *v* on item *i* can be represented by a general linear model

$$S_{vi} = \mu + \pi_v + \beta_i + (\pi\beta,e)_{vi} \ ; \qquad [32]$$

where

S_{vi} = observed score for person *v* on item *i*;
μ = grand mean in the population of persons and the universe of items;
π_v = effect for person *v*;
β_i = effect for item *i*; and
$(\pi\beta, e)_{vi}$ = effect for the interaction of person *v* and item *i*, which is confounded with residual error;

and all effects are assumed to be independent random effects. In the usual case, where each examinee responds once to each item, the interaction effect and the residual error are completely confounded; these two effects are therefore combined in Equation 32. Note that here S_{vi} is an observed score. (Earlier, S_{vk} and S_{wl} were used to denote categories to which persons were assigned.)

The observed score for person *v* is considered to be the mean score over the sample of *n* items. To be consistent with the earlier notation, let the subscript *I* indicate a particular sample of *n* items and let a person's observed mean score be designated:

$$S_{vI} = \mu + \pi_v + \beta_I + (\pi\beta,e)_{vI} \ . \qquad [33]$$

Domain-Referenced Agreement Function

When mastery testing is used to make placement decisions, errors may involve very different degrees of loss. For example, if a mastery test has a cutoff of 80%, the consequences of misclassifying a student with a universe score of 79% may be far less serious than the consequences of misclassifying a student with a universe score of 40%.

This suggests consideration of an agreement function (for domain-referenced tests used for mastery decisions) that involves the distance of the observed score from the cutting score. For a cutting score, λ, a *domain-referenced agreement function* can be defined as

$$d(S_{vI}, S_{wJ}) = (S_{vI} - \lambda)(S_{wJ} - \lambda) \ , \qquad [34]$$

where *I* and *J* refer to independent samples of *n* items. Equation 34 assigns a positive agreement to two scores that result in the same classification, mastery or nonmastery. It assigns a negative agreement to two scores that result in different classifications. In either case, the magnitude of the agreement depends on the magnitudes of two deviation scores, $(S_{vI} - \lambda)$ and $(S_{wJ} - \lambda)$. If both of these deviation scores are close to zero, indicating a "borderline case," the magnitude of the agreement function will be close to zero. If both of these deviation scores are large and in the same direction, indicating strong agreement, the domain-referenced agreement function will be large and positive. If both deviations are large and in opposite directions, indicating strong disagreement, the domain-referenced agreement function will be large and negative.

The agreement function in Equation 34 is similar to the definition of agreement used by Livingston (1972a) in developing an index of reliability for mastery tests. However, Livingston assumed that the two tests were parallel in the sense of classical test theory. The present analysis is based on generalizability theory, which makes the weaker assumption that the tests are randomly parallel. As a result, the indices derived here differ from Livingston's coefficient in several significant ways.

The Index $\theta(d)$

Using the domain-referenced agreement function in Equation 34 and the definition of expected agreement in Equation 2 gives

$$A(d) = \mathop{\mathcal{E}}_{v,I,J} [(S_{vI} - \lambda)(S_{vJ} - \lambda)] \ . \qquad [35]$$

Now, using Equation 33 to replace S_{vI} and S_{vJ} in Equation 35, and noting that the expected value of each cross-product is zero, it follows that

$$A(d) = \mathcal{E}(\mu - \lambda)^2 + \mathcal{E}\pi_v{}^2 + \mathcal{E}\beta_I\beta_J + \mathcal{E}(\pi\beta,e)_{vI}(\pi\beta,e)_{vJ} \ . \qquad [36]$$

Because the two sets of items are independently sampled, the last two terms in Equation 36 equal zero. Also, by the definition of a variance component $\sigma^2(\pi) = \varepsilon\pi_v{}^2$; therefore, the expected agreement for the domain-referenced agreement function is

$$A(d) = (\mu - \lambda)^2 + \sigma^2(\pi) \ . \qquad [37]$$

Similarly, the maximum expected agreement is found by using Equation 34 and the definition of maximum expected agreement in Equation 8a:

$$A_m(d) = \mathop{\mathcal{E}}_{v,I} (S_{vI} - \lambda)^2 \qquad [38a]$$

$$= (\mu - \lambda)^2 + \sigma^2(\pi) + \sigma^2(\beta)/n + \sigma^2(\pi\beta,e)/n \ . \qquad [38b]$$

Substituting Equations 37 and 38b in Equation 7, the index of dependability for mastery decisions is given by

$$\theta(d) = \frac{(\mu - \lambda)^2 + \sigma^2(\pi)}{(\mu - \lambda)^2 + \sigma^2(\pi) + \sigma^2(\beta)/n + \sigma^2(\pi\beta,e)/n} \ . \qquad [39]$$

Equations for estimating $\theta(d)$ have been discussed by Brennan and Kane (1977a) and Brennan (in press). The constant n appears in Equations 38b and 39 because the observed scores are assumed to be averages over n items.

It is clear from Equation 39 that $\theta(d)$ will tend to be large when $(\mu - \lambda)^2$ is large (i.e., when the population mean is very different from the cutting score) even if $\sigma^2(\pi)$ is zero. If all examinees have the same universe score, $\sigma^2(\pi)$ is zero and $(\mu - \lambda)^2$ provides a measure of the strength of the signal that needs to be detected for accurate classification (see Brennan & Kane, 1977b). If this signal is large, the required decisions are easy to make, and it is possible in such cases to classify examinees dependably, even if the test being used does not provide dependable information about individual differences among universe scores.

The Index θ_c(d)

Using the domain-referenced agreement function in Equation 34 and the definition of chance agreement in Equation 15, the expected agreement due to chance is

$$A_c(d) = \mathop{\mathcal{E}}_{v,w,I,J} [(S_{vI} - \lambda)(S_{wJ} - \lambda)] \ . \qquad [40]$$

Replacing S_{vI} and S_{wJ} (see Equation 33) and taking the expected value over v, w, I, and J gives the expected chance agreement for the domain-referenced agreement function:

$$A_c(d) = (\mu - \lambda)^2 . \qquad [41]$$

Subtracting $A_c(d)$ from the numerator and denominator of Equation 39 gives the domain-referenced index of dependability, corrected for chance agreement:

$$\theta_c(d) = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\beta)/n + \sigma^2(\pi\beta,e)/n} \ . \qquad [42]$$

The estimation of this index is discussed by Brennan and Kane (1977b) and Brennan (in press), and its relationship to KR-21 is discussed by Brennan (1977b).

Note that $\theta_c(d)$ is zero when $\sigma^2(\pi)$ is zero. If the test is to provide more dependable classification of examinees than could be achieved by chance, it must differentiate among the examinees. Therefore, some variability in universe scores is required if the test is to make a contribution to the dependability of the decision procedure.

Domain-Referenced Loss and $\sigma^2(\Delta)$

For the domain-referenced agreement function, the expected loss can be found by subtracting Equation 37 from Equation 38b:

$$L(d) = A_m(d) - A(d) = \sigma^2(\beta)/n + \sigma^2(\pi\beta,e)/n. \qquad [43]$$

The loss $L(d)$ is therefore equal to the error $\sigma^2(\Delta)$, which is discussed by Cronbach et al. (1972), Brennan (1977a, 1977b), and Brennan and Kane (1977a, 1977b). The error variance $\sigma^2(\Delta)$ is appropriate for domain-referenced testing, in general, and for mastery testing, in particular.

In mastery testing, interest is in "the degree to which the student has attained criterion performance" (Glaser, 1963, p. 519), independent of the performance of other students. That is, interest is *not* primarily in the relative ordering of examinees' universe scores; rather, it is in the difference between each examinee's universe score and the absolute standard defined by the mastery cutting score. In generalizability theory, the universe score for examinee v is, by definition,

$$\mu_v = \mathop{\mathcal{E}}_{I} S_{vI} = \mu + \pi_v \ , \qquad [44]$$

where S_{vI} is defined by Equation 33, and the expectation is taken over all possible random samples of n items from the universe of items. Therefore, for a mastery test, the error for examinee v is:

$$\Delta_v = (S_{vI} - \lambda) - (\mu_v - \lambda) = \beta_I + (\pi\beta,e)_{vI} \ ; \qquad [45]$$

and the variance of Δ_v over persons and random samples of n items is $\sigma^2(\Delta)$, which is identical to $L(d)$ in Equation 43.

When all students receive the same items, as implied by the linear model in Equation 33, the main effect due to the sampling of items, β_I, affects all examinees' observed scores in the same way. For mastery testing, however, this does not eliminate the item effect as a source of error, because interest is in the *absolute* magnitude of an examinee's score, not the magnitude *relative* to the scores of other examinees. For example, if an especially easy set of items happens to be selected from the universe, the estimates of μ_v (for the universe of items) will tend to be too high for all examinees; this error is accounted for by β_I.

Interpretation of θ(d) and θ_c(d)

The second block of Table 1 summarizes results for the parameters, $A(d)$, $A_m(d)$, $A_c(d)$, and $L(d)$ and the agreement indices $\Theta(d)$ and $\Theta_c(d)$ for the domain-referenced agreement function, d.

The difference in interpretation between $\theta(d)$ and $\theta_c(d)$ parallels the difference between $\theta(t)$ and $\theta_c(t)$. The index $\theta(d)$ characterizes the dependability of decisions or estimates based on the testing procedures. The index $\theta_c(d)$ indicates the *contribution* of the testing procedures to the dependability of these decisions or estimates. It is clear from Equation 39 that $\theta(d)$ may be large even when there is little or no universe score variability in the population of examinees. From Equation 42, however, it can be seen that $\theta_c(d)$ is equal to zero when there is no universe score variability in the population (assuming $\sigma^2(\Delta) > 0$).

Norm-referenced tests compare each examinee's score to the scores of other examinees and therefore require variability if these comparisons are to be dependable. In their now classic paper, Popham and Husek (1969) maintained that "variability is not a necessary condition for a good criterion-referenced test" (p. 3). They argued that since criterion-referenced tests are "used to ascertain an individual's status with respect to some criterion" (p. 2), the meaning of the score is not dependent on comparison with other scores. Popham and Husek conclude, therefore, that indices of dependability that require variability are appropriate for norm-referenced tests but not for criterion-referenced tests.

Although the position adopted by Popham and Husek seems plausible, it leads to a very disturbing conclusion. As Woodson (1974a, p. 64) has pointed out, "items and tests which give no variability . . . give no information and are therefore not useful." This presents, therefore, the paradox that tests providing no information about differences among individual examinees can be good criterion-referenced tests. In two subsequent articles, Millman and Popham (1974) and Woodson (1974b) clarified the two sides of this dispute without resolving the basic issue.

The general framework developed here suggests a resolution of this paradox. As has been seen, two types of coefficients can be developed for any agreement function. Coefficients that are *not* corrected for chance, such as $\theta(d)$, provide estimates of the dependability of the decision procedure; and such coefficients may be large even without variability in universe scores. By contrast, coefficients that *are* corrected for chance, such as $\theta_c(d)$, provide an estimate of the *contribution* of the test to the dependability of the decision procedure. Such coefficients will approach zero as the universe score variance approaches zero. Popham and Husek's argument applies to the decision procedure, and coefficients not corrected for chance are appropriate for characterizing the dependability of the decision procedure. Woodson's argument applies to the *contribution* of the test to the decision procedure, and coefficients corrected for chance are appropriate for characterizing the *contribution* of the test to the dependability of the decision procedure.

Domain-Referenced Agreement Without a Cutting Score

The domain-referenced agreement function in Equation 34 is the product of deviations from a constant. The discussion up to this point has focused on mastery testing, and λ has been considered as the mastery cutting score. However, a single domain-referenced test might be used for several different decisions, involving different cutting scores. In such cases, it would be useful to have an index of dependability that does not depend on a particular cutting score. As discussed earlier, $\theta_c(d)$ is independent of λ and $\theta_c(d)$ is appropriate for assessing the contribution made by the test to the dependability of mastery decisions using any cutting score. Furthermore, $\theta_c(d)$ is less than or equal to $\theta(d)$ for all values of λ, and the two are equal only when $\lambda = \mu$. Therefore, $\theta_c(d)$ provides a lower bound for $\theta(d)$ (see Brennan, 1977b).

Moreover, domain-referenced tests do not necessarily involve any consideration of cutting scores. For example, the score S_{vI} on a domain-referenced test can be interpreted as a descriptive statistic that estimates μ_v, the examinee's universe score (i.e., proportion of items that could be answered correctly) in the domain (see Millman & Popham, 1974). When using domain-referenced scores as descriptive statistics, interest is in point estimates of the examinee's universe score, μ_v. As has been seen, the error (or noise) in such point estimates of universe scores is given by Δ_v, and $\theta_c(d)$ therefore incorporates the appropriate error variance, $\sigma^2(\Delta)$.

The universe score variance, $\sigma^2(\pi)$, in $\theta(d)$ provides a measure of the dispersion of universe scores in the population. There is a strong precedent in physical measurement for taking the variability in universe scores as a measure of the magnitude of the signal to be detected. General purpose instruments for measuring length, for example, are typically evaluated by their ability to detect differences of the order of magnitude of those encountered in some area of practice. Thus, rulers are adequate in carpentry, but verniers are necessary in machine shops.

θ and θ_c for Norm-Referenced Agreement

The agreement function that is implicit in generalizability coefficients (see Cronbach et al., 1972) is

$$g(S_{vI}, S_{wJ}) = (S_{vI} - \mu_I)(S_{wJ} - \mu_J) \quad , \qquad [46]$$

where μ_I is the expected value of S_{vI} over the population of persons for the set of items I; that is,

$$\mu_I = \mathcal{E}_v S_{vI} = \mu + \beta_I \; . \qquad [47]$$

The parameter β_I is the average value of the item effect for the I^{th} sample of n items.

Note that the agreement function for norm-referenced tests, given in Equation 46, and the agreement function for domain-referenced tests, given in Equation 34, are both products of deviation scores. The difference between the two agreement functions is in the nature of the deviation scores that are used. The norm-referenced agreement function is defined in terms of deviations from the population mean for fixed sets of items. These deviation scores compare the examinee's performance on the set of items to the performance of the population on the same set of items. The domain-referenced agreement function in Equation 34 is defined in terms of the deviations of the examinees' scores from a fixed cutting score.

The Indices $\theta(g)$ and $\theta_c(g)$

Using Equations 33, 46, and 47, the expected norm-referenced agreement is

$$A(g) = \mathcal{E}_{v,I,J} [(S_{vI} - \mu_I)\ (S_{vJ} - \mu_J)] = \sigma^2(\pi) \; . \qquad [48]$$

Also, the maximum expected agreement for the norm-referenced agreement function is

$$A_m(g) = \mathcal{E}_{v,I} (S_{vI} - \mu_I)^2 = \sigma^2(\pi) + \sigma^2(\pi\beta,e)/n \; . \qquad [49]$$

Substituting Equations 48 and 49 in Equation 7, an index of dependability for norm-referenced tests is

$$\theta(g) = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\pi\beta,e)/n} \; . \qquad [50]$$

Using the norm-referenced agreement function in Equation 46 and the definition of chance agreement in Equation 15, the expected agreement due to chance is zero; and therefore

$$\theta_c(g) = \theta(g) \; . \qquad [51]$$

The correction for chance agreement has no effect on the norm-referenced dependability index because it has a built-in correction for chance.

Norm-Referenced Loss and $\sigma^2(\delta)$

The loss associated with the norm-referenced agreement function is found by subtracting Equation 48 from 49:

$$L(g) = A_m(g) - A(g) = \sigma^2(\pi\beta,e)/n \; . \qquad [52]$$

This loss is simply the error variance designated by Cronbach et al. (1972) as $\sigma^2(\delta)$, which is also the error variance in classical test theory.

In norm-referenced testing, interest is in "the relative ordering of individuals with respect to their test performance, for example, whether student A can solve his [or her] problems more quickly than student B" (Glaser, 1963, p. 519). Thus, interest is in "the adequacy of the measuring procedure for making *comparative* decisions" (Cronbach et al., 1972, p. 95). In this situation the error for a given person, as defined by Cronbach et al. (1972) is

$$\delta_v = (S_{vI} - \mu_I) - (\mu_v - \mu) \qquad [53a]$$

$$= [\mu + \pi_v + \beta_I + (\pi\beta,e)_{vI} - \mu - \beta_I] - [\mu + \pi_v - \mu] \qquad [53b]$$

$$= (\pi\beta,e)_{vI} ; \qquad [53c]$$

and the variance of δ_v over the population of persons and samples of n items is

$$\sigma^2(\delta) = \sigma^2(\pi\beta,e)/n = L(g) . \qquad [54]$$

Substituting Equations 48 and 54 in Equation 22 gives

$$\theta(g) = \theta_c(g) = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\delta)} , \qquad [55]$$

which is identical to the generalizability coefficient $\varepsilon\rho^2$, given the random effects linear model in Equation 33. (Equation 55 is also equivalent to Cronbach's (1951) coefficient alpha and to KR-20 for dichotomously scored items.)

Interpretation of $\theta(g) = \theta_c(g)$

The third block of Table 1 summarizes results for the parameters $A(g)$, $A_m(g)$, $A_c(g)$, and $L(g)$ and the agreement indices $\theta(g)$ and $\theta_c(g)$ for the norm-referenced agreement function, g. Note that $\theta_c(d)$ and $\theta_c(g)$ incorporate the same expected agreement (or signal) but different definitions of error variance (loss or noise). For $\theta_c(d)$ the error variance is $\sigma^2(\Delta)$ and for $\theta_c(g)$ the error variance is $\sigma^2(\delta)$. Since

$$\sigma^2(\delta) \leq \sigma^2(\Delta) , \qquad [56]$$

it follows that

$$\theta_c(d) \leq \theta_c(g) . \qquad [57]$$

The difference between $\sigma^2(\Delta)$ and $\sigma^2(\delta)$ is simply $\sigma^2(\beta)/n$. Therefore, $\theta_c(d)$ and $\theta_c(g)$ are equal only when β_I is a constant for all instances of the testing procedure. The variance component for the main effect for items, $\sigma^2(\beta)$, reflects differences in the mean score (in the population) for different samples of items. If interest is only in differences among examinee universe scores, as in norm-referenced testing, then any effect that is a constant for all examinees does not contribute to the error variance. However, for domain-referenced testing, interest is in the absolute magnitude of examinee universe scores, or the magnitude compared to some externally defined cutting score. In this case, fluctuations in mean scores for samples of items *do* contribute to error variance.

The Effect of Item Sampling on the Indices of Dependability, θ and θ_c

Items Nested Within Persons in the D Study

The implications of using several definitions of agreement for randomly parallel tests have been examined. It has been assumed that for each instance of the testing procedure a random sample of items from some infinite domain is administered to all examinees, i.e., items are crossed with examinees. Following Cronbach et al. (1972), this design is designated $p \times i$. Indices that are appropriate for other designs can be derived using the same approach. A particularly interesting and useful set of indices is obtained by assuming that an independent random sample of items is selected for *each* examinee. Following Cronbach et al. (1972), this design is designated *i:p*, where the colon means "nested within."

In this section it will be convenient to make use of the distinction between a G study and a D study—a distinction originally drawn by Rajaratnam (1960) and subsequently discussed extensively by Cronbach et al. (1972). The purpose of a G study, or generalizability study, is to examine the dependability of some measurement procedure. The purpose of a D study, or decision study, is to provide the data for making substantive decisions. "For example, the published estimates of reliability for a college aptitude test are based on a G study. College personnel officers employ these estimates to judge the accuracy of data they collect on their applicants (D study)" (Cronbach et al., 1972, p. 16). The G study generates estimates of variance components, which can then be used in a variety of D studies. Generally, G studies are most useful when they employ crossed designs and large sample sizes to provide stable estimates of as many variance components as possible.

In previous sections of this paper, it has been implicitly assumed that both the G study and the D study used the crossed design, $p \times i$. It will continue to be assumed that variance components have been estimated from the crossed design. However, in this section it will be assumed that the D study employs the *i:p* design. For example, in computer-assisted testing it is frequently desirable (or even necessary for security reasons) that each examinee receive a different set of items, i.e., the D study uses an *i:p* design. However, even in such cases, it is desirable that the variance components be estimates from the crossed design, $p \times i$.

If in the D study each examinee gets a different set of items, the item effect will not be the same for all examinees. Under these circumstances, the linear model for scores on a particular instance of the testing procedure is

$$S_{vI} = \mu + \pi_v + \beta_{vI} + (\pi\beta,e)_{vI} \; ; \qquad [58]$$

and the item effect is now confounded with the residual, $(\pi\,\beta, e)_{vI}$. It is particularly important to note that for Equation 58, the expected value of S_{vI} over persons (v) is μ, where μ is the grand mean in the population of persons and the universe of items. The population mean of the observed scores, S_{vI}, does *not* equal μ_I, which is the expected value over the population for a particular set of items, I. When items are nested within persons, taking the expected value of the observed scores over the infinite population of examinees implies taking the expected value over an infinite universe of items.

Implications for Norm-Referenced and Domain-Referenced Indices of Dependability

Using the linear model in Equation 58 and the norm-referenced agreement function in Equation 46, it can be shown that

$$A(g') = \sigma^2(\pi) \; ; \qquad [59]$$

$$A_m(g') = \sigma^2(\pi) + \sigma^2(\beta)/n + \sigma^2(\pi\beta,e)/n \; ; \text{ and} \qquad [60]$$

$$A_c(g') = 0 \; ; \qquad [61]$$

where the prime following *g* differentiates quantities associated with the nested design, *i:p*, from quantities associated with the crossed design, *p* × *i*. Substituting these results in Equations 7 and 19 gives

$$\theta(g') = \theta_c(g') = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + [\sigma^2(\beta) + \sigma^2(\pi\beta,e)]/n} \qquad [62a]$$

$$= \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\Delta)} \; . \qquad [62b]$$

Note that both $\theta(g')$ and $\theta_c(g')$ are identical to $\theta_c(d)$, the domain-referenced dependability index corrected for chance in Equation 42. The only difference between $\theta(g')$ and the usual dependability index for norm-referenced tests, $\theta(g)$, is that $\theta(g')$ has an additional term, $\sigma^2(\beta)/n$, in the denominator.

For norm-referenced tests, when the same items are administered to all examinees, the item effect, β_I, is a constant for all examinees and $\sigma^2(\beta)/n$ does *not* enter the error variance. If items are nested within examinees, however, β_{vI} will generally be different for each examinee and $\sigma^2(\beta)/n$ is part of the error variance.

For the domain-referenced agreement function, the agreement indices developed from the nested model are identical to those developed from the crossed model. The dependability of a domain-referenced testing procedure is not affected by whether the D study uses the crossed design, *p* × *i*, or the nested design, *i:p*. The aim of domain-referenced testing is to provide point estimates of examinee universe scores rather than to make comparisons among examinees. The dependability of each examinee's score is determined by the number of items administered to that examinee, not by how many items or which items are administered to other examinees.

Standardization of the items used in any instance of the testing procedure improves the dependability of norm-referenced interpretations but does not improve the dependability of domain-referenced interpretations. Furthermore, the use of different samples of items for different examinees will tend to improve estimates of group means. Therefore, if domain-referenced tests are to be used for program evaluation, the selection of independent samples of items for different examinees provides *more* dependable estimates of group means without any loss in the dependability of estimates of examinees' universe scores. Lord (1977) arrives at a similar conclusion based on his consideration of computer-generated repeatable tests.

Summary and Conclusions

Table 1 provides an overview of the major results derived and discussed in this paper. Two indices of dependability, θ and θ_c, were discussed for three different agreement functions: the threshold agreement function, t; the domain-referenced agreement function, d; and the norm-referenced agreement function, g. This paper emphasized considerations relevant to the first two agreement functions because the indices of dependability associated with them are indices that have been proposed for domain-referenced and mastery tests. The norm-referenced agreement function, g, was considered primarily for purposes of comparing it with the other two agreement functions. The main purposes of this paper were to provide an internally consistent framework for deriving indices of dependability for domain-referenced tests and to examine the implications of choosing a particular index.

Choosing an Index of Dependability

Although the discussion of these issues has not dictated which index should be chosen in a particular context, it has indicated that two main issues are involved in such a choice: (1) the nature of agreement functions (or, alternatively, loss functions) and (2) the use of an index corrected for chance or not corrected for chance.

With respect to the first issue, two types of agreement functions have been considered for mastery tests: the threshold agreement function and the domain-referenced agreement function. The threshold agreement function is appropriate whenever the main distinction that can be made usefully is a *qualitative* distinction between masters and nonmasters. If, however, different degrees of mastery and nonmastery exist to an appreciable extent, the threshold agreement function may not be appropriate, since it ignores such differences.

In many educational contexts, differences between masters and nonmasters are not purely qualitative. Rather, the attribute that is measured is conceptualized as an ordinal or interval scale; and the examinees may possess the attribute to varying degrees, even though a single cutting score is used to define mastery. In this context the misclassification of examinees who are far above or below the cutting score is likely to cause serious losses. The misclassification of examinees whose level of ability is close to the cutting score will usually involve much less serious losses. The domain-referenced agreement function, d, assigns a positive value to the agreement whenever both instances of the testing procedure place the examinee in the same category, and it assigns a negative value to the agreement when the two instances place an examinee in different categories. Furthermore, the magnitude of the agreement is determined by the distance of the observed scores from the cutting score on the two tests.

The second issue in choosing an index of dependability is whether to use an index θ, which is not corrected for chance agreement, or an index θ_c, which *is* corrected for chance. There is no reason to prefer one index over the other in all contexts. The two indices provide different information and therefore should be interpreted differently. For judgments about the dependability of a decision procedure as applied to a particular population, indices that are *not* corrected for chance are appropriate. For judgments about the *contribution* of tests to the dependability of the decision procedure, indices that are corrected for chance are appropriate. (It is to be recognized, of course, that the word "chance" is used in the specific sense indicated by Equations 15, 16, or 18.)

It is also useful to note that whether θ or θ_c is chosen, the expected loss or error variance remains unchanged. That is, the choice between θ and θ_c usually affects the strength of the signal in a testing procedure but never the strength of the noise (see Brennan & Kane, 1977b). In effect, when θ_c is chosen, the strength of the signal is reduced by an amount attributable to chance, and it is this reduction of signal strength that usually causes θ_c to be less than θ. As noted previously, for the norm-referenced agreement function, g, θ always equals θ_c because chance agreement is zero. Indeed, this is probably one reason why the distinction between indices such as θ and θ_c has been ignored in much of the literature on testing and psychometrics.

Prior Information

For the domain-referenced agreement function, d, $\theta(d)$ equals $\theta_c(d)$ when $(\mu - \lambda)^2$ equals zero, i.e., when the mean, μ, equals the cutting score, λ. In such cases, prior information about μ is of no use in classifying examinees as masters or nonmasters; and the dependability of decisions depends entirely upon the dependability of the test being used. If $(\mu - \lambda)^2$ is very large, decisions made about a student's mastery or nonmastery status, solely on the basis of prior information about μ, may be highly dependable. If, however, $(\mu - \lambda)^2$ is nonzero, but not very large compared to the expected loss, $\sigma^2(\Delta)$, it is likely that the dependability of decisions could be improved by using Bayesian methods.

Bayesian procedures (Hambleton & Novick, 1972; Swaminathan, Hambleton, & Algina, 1975) take advantage of prior information about the population by using this information and the student's observed score to estimate the student's universe score. The optimum weighting of prior information and test scores depends on the prior distribution of universe scores in the population, the dependability of the testing procedure, and the agreement function (or equivalently, the loss function) that is chosen. Although most recent published applications of Bayesian methods in domain-referenced testing have used threshold loss, Bayesian methods are in principle equally applicable for squared error loss.

Assumptions About Parallel Tests

Throughout this paper, it has been assumed that two tests are parallel if they involve random samples of the same number of items from the same universe, or domain, of items. That is, the assumption of randomly parallel tests, rather than the *stronger* assumption of classically parallel tests, has been made. Cronbach et al. (1972) have shown that either assumption can be used as a basis for defining the generalizability coefficient for the persons-crossed-with-items design; and it has been shown that this generalizability coefficient is identical to $\theta(g) = \theta_c(g)$ for norm-referenced tests. Also, either assumption, in conjunction with the threshold agreement function, can be used to derive the indices $\theta(t)$ and $\theta_c(t)$.

It seems that the assumption of classically parallel tests is usually too restrictive for a domain-referenced test. However, if all items in the universe are equally difficult for the population of persons, then the item effect, β_i, is a constant for all items and $\sigma^2(\Delta)$ equals $\sigma^2(\delta)$. That is, the expected loss for the domain-referenced agreement function equals the expected loss for the norm-referenced agreement function. In this case the index $\theta(d)$ is identical to Livingston's (1972a, 1972b, 1972c, 1973) coefficient.

The differences between $\theta(d)$ and Livingston's coefficient are therefore a direct result of the differences between the assumptions of randomly parallel tests and classically parallel tests, respectively. It is important to note, however, that neither index is corrected for chance. They both reflect the dependability of a decision procedure, *not* the contribution of tests to the dependability of a decision procedure. Also, for both coefficients, changes in the cutting score, λ, affect the coefficients' magnitudes through the signal strength, *not* through the noise or error variance.

Concluding Comments

This paper has concentrated on indices of dependability for domain-referenced tests and factors that influence the use and interpretation of such indices. Particular emphasis has been on the indices $\theta(d)$ and $\theta_c(d)$ because (1) they have broad applicability in domain-referenced testing, (2) they are easily compared with the usual norm-referenced indices of dependability, and (3) they can be developed using principles from generalizability theory, a broadly applicable psychometric model. Using principles from generalizability theory, it is relatively straightforward to define $\theta(d)$ and $\theta_c(d)$ for ANOVA designs other than the simple persons-crossed-with-items design. The extension of $\theta(t)$ and $\theta_c(t)$ to other designs is not so straightforward.

However, regardless of which index of dependability an evaluator chooses, it is important that the evaluator recognize the underlying assumptions and interpret results in a meaningful manner. In this regard, it is often the case that the magnitude of an index of dependability, alone, provides an insufficient basis for decision making. It is almost always best to provide, also, the quantities that enter the index (A, A_m, A_c, and L in Table 1), as well s the estimated variance components (See American Psychological Association, 1974).

References

American Psychological Association. *Standards for educational and psychological tests* (Rev. ed.). Washington, DC: Author, 1974.

Brennan, R. L. *Generalizability analyses: Principles and procedures* (ACT Technical Bulletin No. 26). Iowa City, IA: The American College Testing Program, September 1977. (a)

Brennan, R. L. *KR-21 and lower limits of an index of dependability for mastery tests* (ACT Technical Bulletin No. 27). Iowa City, IA: The American College Testing Program, December 1977. (b)

Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art.* Baltimore, MD: The Johns Hopkins University Press, in press.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. *Journal of Educational Measurement,* 1977, *14,* 277-289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. *Psychometrika,* 1977, *42,* 609-625. (Errata. *Psychometrika,* 1978, *43,* 289.) (b)

Carver, R. P. Special problems in measuring change with psychometric devices. In *Evaluative research: Strategies and methods.* Pittsburgh, PA: American Institutes for Research, 1970.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 1960, *20,* 37-46.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika,* 1951, *16,* 292-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 1963, *18,* 519-521.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education, 1971.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement,* 1973, *10,* 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research,* 1978, *48,* 1-47.

Hively, W. Introduction to domain-referenced testing. In W. Hively (Ed.), *Domain-referenced testing.* Englewood Cliffs, NJ: Educational Technology Publications, 1974.

Huynh, H. On consistency of decisions in criterion-referenced testing. *Journal of Educational Measurement,* 1976, *13,* 265-275.

Livingston, S. A. A criterion-referenced application of classical test theory. *Journal of Educational Measurement,* 1972, *9,* 13-26. (a)

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." *Journal of Educational Measurement,* 1972, *9,* 31. (b)

Livingston, S. A. Reply to Shavelson, Block, and Ravitch's "Criterion-referenced testing: Comments on reliability." *Journal of Educational Measurement,* 1972, *9,* 139. (c)

Livingston, S. A. A note on the interpretation of the criterion-referenced reliability coefficient. *Journal of Educational Measurement,* 1973, *4,* 311.

Lord, F. M. Some item analysis and test theory for a system of computer-assisted test construction for individualized instruction. *Applied Psychological Measurement,* 1977, *1,* 447-455.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education.* Berkeley, CA: McCutchan, 1974.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. *Journal of Educational Measurement,* 1974, *11,* 137-138.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement,* 1969, *6,* 1-9.

Rajaratnam, N. Reliability formulas for independent decision data when reliability data are matched. *Psychometrika,* 1960, *25,* 261-271.

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly,* 1955, *19,* 321-325.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement,* 1976, *13,* 253-264.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement,* 1978, *15,* 111-116.

Subkoviak, M. J. Decision-consistency approaches. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art.* Baltimore, MD: The Johns Hopkins University Press, in press.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision theoretic formulation. *Journal of Educational Measurement,* 1974, *11,* 263-267.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement,* 1975, *12,* 87-98.

Woodson, M. I. The issue of item and test variance for criterion-referenced tests. *Journal of Educational Measurement,* 1974, *11,* 63-64. (a)

Woodson, M. I. The issue of item and test variance for criterion-referenced tests: A reply. *Journal of Educational Measurement,* 1974, *11,* 139-140. (b)

Acknowledgments

This research was partially supported by Contract No. N00123-77-C-0739 between the American College Testing Program and the Navy Personnel Research and Development Center.

Author's Address

Send requests for reprints or further information to Michael T. Kane, Director of Test Development, National League for Nursing, 10 Columbus Circle, New York, NY 10019.