ACT Research Report Series 2001-1

Suggestions for the Evaluation and Use of Concordance Results

Bradley A. Hanson
Deborah J. Harris
Mary Pommerich
James A. Sconing
Qing Yi

ACT

February 2001

Suggestions for the Evaluation and Use of Concordance Results

Bradley A. Hanson, Deborah J. Harris, Mary Pommerich, James A. Sconing, Qing Yi
ACT, Inc.

Abstract

A linkage between scores on two tests that do not measure the same underlying construct is called a concordance. This paper discusses the evaluation and appropriate uses of concordances. A conceptualization of score equivalence in terms of a latent variable model for test scores is presented. Two factors involved in evaluating the quality of a linkage of two scores are discussed: 1) the initial comparability of the scores, and 2) whether comparability is desired for scores of individual examinees, or for score distributions. It is concluded that the only appropriate use of concordant scores of individuals is in situations where empirical evidence exists that the inferences made using concordant scores and the inferences made using the actual scores are not too different. The use of concordant score distributions is likely to be less problematic than the use of concordant scores of individuals. Still, evidence should exist that a concordant score distribution does not differ greatly from the distribution of the actual scores in any population in which the concordance is used. Examples are presented to illustrate the evaluation and appropriate uses of concordances.

Acknowledgements

The authors thank Richard Sawyer, Dave Woodruff, and Mike Kolen for reviewing an earlier version of this paper and providing helpful comments.

Suggestions for the Evaluation and Use of Concordance Results

The focus of this paper is on issues involved in evaluating the quality of a concordance, and the appropriate uses of a concordance. The first section discusses concepts of score equivalence in the context of a latent variable model for test scores. These concepts are used to distinguish among various types of linkages, including concordance, and are useful in considering the quality of a linkage. In the second section the concepts developed in the first section are used to discuss the evaluation of concordances, and the appropriate uses of concordances.

Score Equivalence

In this paper the term *test* will denote a set of specifications that describe how to build test forms. A test form consists of a specific set of items that meet a set of test specifications. Test specifications include information about the type and content of items on a test form, as well as administration conditions. The development of test specifications is part of a process that involves the more general steps of defining a domain and a framework of particular content within the domain to be assessed (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999).

This paper deals with the case in which the responses of examinees to items on a particular test form are translated into a discrete univariate test score. A linkage between two scores is a function that transforms one score (denoted Y) to attempt to make it comparable with the other score (denoted X). A linking function is applied to values of score Y obtained by individual examinees with the intention of using the transformed score Y as if it were score X. The linking function can also be applied to a distribution of Y scores with the intention of using the transformed score Y distribution as if it were a score X distribution. This paper only considers linkages for observable test scores. Linkages of latent variable distributions (e.g., Bloxom, Pashley, Nicewander, & Yan, 1995; Williams, Rosa, McLeod, Thissen, & Sanford, 1998) are not considered.

An important factor in evaluating the quality of a linkage is the initial comparability of the two scores for which a linkage is sought. Whether two scores are measuring the same underlying variables (measuring the same thing) along with the concepts of first and second order equity (Lord, 1980; Morris, 1982) are used in this paper as a means of evaluating the comparability of two scores. At one extreme two scores can be measuring the same thing and in addition be close to achieving first and second order equity as defined below. Scores

that meet these conditions will be denoted *closely equable* scores. At the other extreme are scores that are not measuring the same thing, which will be denoted *nonequable scores.* In between these two extremes are *weakly equable* scores that are measuring the same thing, but come from forms that are not designed to be parallel, so it is expected the scores will deviate more from first and/or second order equity than closely equable scores. These three types of score comparability represent three useful reference points on a continuum representing the degree of similarity between the scores to be linked.

The established terminology used to describe the linking of closely equable and weakly equable scores are equating and calibration, respectively (Linn, 1993; Mislevy, 1992; Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Kolen & Brennan, 1995). The terminology used to refer to linking nonequable scores depends on the method used to compute the linking function. When regression is used to compute the linkage function the resulting linkage has been termed prediction (Linn, 1993; Mislevy, 1992; Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). When equipercentile methods are used to compute the linkage function the linkage is termed a concordance (Marco & Abdel-Fattah, 1991; Houston & Sawyer, 1991; Dorans, Lyu, Pommerich, & Houston, 1997). When the linkage function is computed using methods involving moderator variables the linkage is termed statistical moderation (Linn, 1993; Mislevy, 1992; Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). Mislevy (1992) also uses the term statistical moderation to refer to concordance (which is called case 1 statistical moderation). In this paper the term concordance is used to indicate a linking of nonequable scores regardless of how the linking function is computed.

Another factor in evaluating the quality of a linkage is the level at which the score comparability is desired. The quality of a linkage can be evaluated at two levels: 1) individual examinee scores, and 2) score distributions. Comparability of individual examinee scores would imply comparability of score distributions, but it is possible a linkage could result in a high degree of comparability of score distributions while not providing a high degree of score comparability for some individuals. For instance, it is always possible to develop a link function that results in almost perfect comparability of *distributions* in *one* population, no matter how incomparable the two scores are for individuals.

The next two subsections discuss the comparability of individual scores and score distributions, respectively.

Comparability of Individual Scores

For two tests X and Y assume there is a latent random vector $\boldsymbol{\Theta}$ that accounts for all

the systematic variation of the scores on the two tests. The variables in the latent random vector $\boldsymbol{\Theta}$ are a union of the latent variables measured by tests X and Y.

Let X be the random variable representing the score on a form of test X, and Y be the random variable representing the score on a form of test Y in a particular population of examinees. The function that gives the true score of X as a function of $\boldsymbol{\theta}$ (a realization of $\boldsymbol{\Theta}$) is $\tau_x(\boldsymbol{\theta}) = E(X \mid \boldsymbol{\theta})$, and the function that gives the true score of Y as a function of $\boldsymbol{\theta}$ is $\tau_y(\boldsymbol{\theta}) = E(Y \mid \boldsymbol{\theta})$. Two scores are said to measure the same thing if the true score on one is a function of the true score on the other (i.e., given one true score the other true score is uniquely determined), otherwise they are said to measure different things.

If $t(Y)$ be an increasing function that transforms score Y in an attempt to make it comparable with score X, then the deviation from equity for scores X and $t(Y)$ (the actual and concordant scores) at x and $\boldsymbol{\theta}$ is:

$$DE(x, \boldsymbol{\theta}) = F_{t(Y)}(x \mid \boldsymbol{\theta}) - F_X(x \mid \boldsymbol{\theta}), \tag{1}$$

where $F_X(x \mid \boldsymbol{\theta})$ is the conditional cumulative distribution function for X given $\boldsymbol{\Theta} = \boldsymbol{\theta}$ ($\Pr[X \leq x \mid \boldsymbol{\theta}]$), and $F_{t(Y)}(x \mid \boldsymbol{\theta})$ is the conditional cumulative distribution function for $t(Y)$ given $\boldsymbol{\Theta} = \boldsymbol{\theta}$ ($\Pr[t(Y) \leq x \mid \boldsymbol{\theta}]$). If $DE(x, \boldsymbol{\theta}) = 0$ for all x and $\boldsymbol{\theta}$ then equity would hold for scores $t(Y)$ and X. Lord (1980) showed that equity cannot be achieved even for scores measuring the same thing unless the two scores are parallel or both scores are perfectly reliable. If equity cannot be achieved for scores measuring the same thing, it also cannot be achieved for scores measuring different things.

By assumption, all the systematic variance in Y and X is accounted for by $\boldsymbol{\Theta}$. Then, in addition to the marginal distributions of $t(Y)$ and X being identical for all $\boldsymbol{\theta}$ if $DE(x, \boldsymbol{\theta}) = 0$ for all x and $\boldsymbol{\theta}$, the associations of X and $t(Y)$ with any other variable will be identical. Consequently, if the deviation from equity is zero for all x and $\boldsymbol{\theta}$, then any inferences using $t(Y)$ would be the same as those using X. An evaluation of the comparability of two scores should focus on how close to zero the deviation from equity is, which indicates the extent to which inferences made using X and $t(Y)$ tend to be the same. The deviation from equity being zero for all x and $\boldsymbol{\theta}$ is an ideal that cannot ever realistically be fully met, not a criterion to be judged as being met or not.

The deviation from equity for the first two moments of the conditional distributions given in Equation 1 (first and second order equity) captures important aspects of score comparability. The deviation from first order equity for scores $t(Y)$ and X at $\boldsymbol{\theta}$ is

$$E[t(Y) \mid \boldsymbol{\theta}] - E[X \mid \boldsymbol{\theta}], \tag{2}$$

where the first expected value is over the conditional distribution of Y given $\boldsymbol{\Theta}$, and the second expected value is over the conditional distribution of X given $\boldsymbol{\Theta}$. The first expected value in Equation 2 is equal to the true score corresponding to the observed score $t(Y)$, and the second expected value is equal to the true score corresponding to the observed score X. Equation 2 is the difference in the true scores corresponding to $t(Y)$ and X. The deviation from second order equity for scores $t(Y)$ and X at $\boldsymbol{\theta}$ is

$$\sigma^2[t(Y) \mid \boldsymbol{\theta}] - \sigma^2[X \mid \boldsymbol{\theta}]\,. \tag{3}$$

Equation 3 gives the difference in the conditional measurement errors for scores $t(Y)$ and X.

Unless the scores to be linked are parallel, no function t exists such that even the deviation from first order equity is zero. Therefore, the comparability of $t(Y)$ and X should focus on how close the deviations from first and second order equity are to zero.

Two cases of score comparability are distinguished when scores X and Y measure the same thing. One case is when the test forms are designed so that the deviations from both first and second order equity for scores X and Y should be small (e.g., scores from two forms produced from the same specifications and designed to be parallel). In this case scores X and Y will be referred to as *closely equable*, the linkage of X and Y is called an equating, and t is called an equating function (Linn, 1993; Mislevy, 1992). In this case t serves to fine tune the score Y so the deviation from equity of $t(Y)$ and X is less than the deviation from equity of Y and X.

Another case occurs when the forms are designed to measure the same thing, but are not designed to be parallel, so it is expected that scores X and Y will deviate more from first and/or second order equity than closely equable scores. For example, scores on two forms produced from the same specifications, but containing different numbers of items, or scores designed to measure the same thing but at different grade levels. The scores X and Y in this case will be referred to as *weakly equable.* A term used to refer to a linkage between weakly equable scores is calibration (Linn, 1993; Mislevy, 1992). The term *calibration* is generally reserved for linkage of weakly equable scores using methodology involving latent variable models. The term *vertical equating* or *vertical scaling* has been used to refer to linking weakly equable scores designed to be used at different grade levels.

The above discussion considers score comparability for two specific forms of two different tests. In many cases a function t is computed that is applied to scores from a variety of forms that are already equated. The data used to develop such a t are typically equated

scores from a group of examinees who have taken a number of different forms of each test. The degree to which the converted score on test Y is comparable to a score on test X will be form-dependent because forms will not be perfectly equated (i.e., deviations from equity will differ for different pairs of forms). For example, the concordance between ACT and SAT is developed using equated scale scores for the two tests. A complete evaluation of the quality of the ACT and SAT concordance in terms of deviations from equity as described in this section is complicated by the fact that the concordance is based on multiple forms of the ACT and SAT, whereas the deviations from equity probably depend on the forms considered.

Comparability of Score Distributions

Let $F_Y(x)$ and $F_X(x)$ be the marginal cumulative distribution functions for Y and X in a particular population (marginalized over the latent vector $\boldsymbol{\Theta}$). While no linking function t can be found that results in equity of the individual scores X and Y, if X and Y are continuous random variables a function t exists such that

$$F_{t(Y)}(x) = F_X(x)\,, \tag{4}$$

for all x. Equation 4 will hold for a function t that transforms Y such that the percentiles of $t(Y)$ and X are equal (the p-th percentile of a continuous random variable X is the value x such that $F_X(x) = p/100$). This function is the equipercentile function given by

$$t_{ep}(y) = F_X[F_Y^{-1}(y)]\,. \tag{5}$$

The definition of the equipercentile function only exists when Y and X are continuous random variables. In the case considered here where test scores X and Y are discrete an equipercentile function can be defined by continuizing X and Y (Holland & Thayer, 1989). Since there is more than one way to continuize X and Y a unique equipercentile function does not exist when the random variables are discrete. The most common way of continuizing X and Y in order to compute an equipercentile function is to spread out the discrete density using a uniform kernel (Holland & Thayer, 1989; Hanson, 1993).

The score given by the equipercentile function is only called an equated score when applied to scores that are closely equable. When the equipercentile function (or any linking function) is used to link two scores that are nonequable the score produced by the equipercentile function would be called a concordant score rather than an equated score. The distinction between the equipercentile function producing an equated versus concordant

score is not relevant from the standpoint of the distribution of scores. In other words, the equipercentile function makes the marginal distributions of two scores identical whether or not the scores are closely equable, weakly equable, or nonequable.

If the equipercentile function is computed using data from a single group design (where a group of examinees takes both tests), or a random groups design (where randomly equivalent groups of examinees each take one of the two tests), then it will only function to convert the distribution of Y to be the same as the distribution of X in the population from which examinees were sampled. In practical settings the equipercentile function may be used to produce a concordance used with individual scores. In this case the distinction between concordant and equated scores is important, as discussed in the previous sections.

The equipercentile function might also be applied to distributions of Y from different populations than that for which the equipercentile function was computed. Let $F_X(x \mid z)$ and $F_Y(y \mid z)$ be cumulative distribution functions for the conditional distributions of X given $Z = z$ and Y given $Z = z$, where Z is a variable which takes on values that are indicators of different populations of interest (e.g., males or females, different states, or different schools). If an equipercentile function is computed for population $Z = z_0$ ($t_0(y) = F_X[F_Y^{-1}(y \mid z_0)]$) then it will generally not be the case that

$$F_{t_0(Y)}(x \mid z) - F_X(x \mid z) = 0\,, \tag{6}$$

for $z \neq z_0$ and all x. The left side of Equation 6 gives the deviation from equity of score distributions, which can be written as

$$\int_S [F_{t_0(Y)}(x \mid \boldsymbol{\theta}, z) - F_X(x \mid \boldsymbol{\theta}, z)]g(\boldsymbol{\theta} \mid z)d\boldsymbol{\theta}\,, \tag{7}$$

where $g(\boldsymbol{\theta} \mid z)$ is the conditional density of $\boldsymbol{\Theta}$ given $Z = z$, and S is the region over which the density of $\boldsymbol{\Theta}$ is non-zero. Equation 7 can be written as

$$\int_S [F_{t_0(Y)}(x \mid \boldsymbol{\theta}) - F_X(x \mid \boldsymbol{\theta})]g(\boldsymbol{\theta} \mid z)d\boldsymbol{\theta}\,, \tag{8}$$

since it is assumed the random vector $\boldsymbol{\Theta}$ accounts for all the systematic variation in Y and X. From Equation 8 it can be seen that the deviation in equity of score distributions will depend on how close the deviation from equity for individual scores in Equation 1 is to zero. The deviation from equity of score distributions can be close to zero as long as the deviation from equity of individual scores is not too far from zero for values of the latent

variables for which the density $g(\boldsymbol{\theta} \mid z)$ is large. Thus, the deviation from equity of score distributions will likely be closer to zero than the deviation from equity of at least some individual scores.

Methods of Computing Linkage Functions

Mislevy (1992) and Linn (1993) have identified five types of linkages: equating, calibration, prediction, statistical moderation, and social moderation. The first two linkage types correspond to linkages for closely equable and weakly equable scores, respectively. The last three types of linkages all refer to linking nonequable scores, and are distinguished by the method used to compute the linking function. For instance, prediction is a type of linking produced when regression is used to compute the linking function. Statistical moderation is a type of linking produced when two scores are statistically linked through a third moderator variable. Judgments about the comparability of performance on two tests are used to create a linking function in social moderation. The approach taken in this paper is to distinguish linkage types by the characteristics of the scores to be linked (closely equable, weakly equable, or nonequable) rather than the method used to compute the linkage function.

Evaluating and Using Concordances

This section discusses appropriate use of linkages for nonequable scores. Linkages for nonequable scores will be referred to as concordances regardless of what method is actually used to compute the linking function. Previous sections made the distinction between applying a linking function to scores for individual examinees versus applying a linking function to a score distribution for some group. The next two sections discuss the evaluation and appropriate use of concordance results for individual scores and score distributions, respectively, using concepts concerning the quality of a concordance described in the previous section.

Evaluation and Appropriate Uses of Concordant Individual Scores

Only if the deviations from first and second order equity given in Equations 2 and 3 are both close to zero will $t(Y)$ and X be called interchangeable in the sense that for a wide range of purposes similar inferences would be made using either X or $t(Y)$. The deviations from first and second order equity are written in terms of a latent variable model. A way to evaluate a concordance using deviations from equity would be to collect data, estimate the parameters of a latent variable model using the data, and compute the differences in Equations 2 and 3 as a function of the latent variables using the parameter estimates.

These results could be complex to evaluate, especially in the case of nonequable scores where equity will depend on a vector of latent variables.

Since zero deviations from first and second order equity are ideals rather than achievable goals, the question of evaluating equity becomes a question of how close to zero the deviation from equity needs to be in order to say that a concordant score can be used interchangeably with the actual score at the individual level. In this paper it is assumed that the deviation from first and second order equity should be close to zero over the latent variable range where most examinees fall in order for two scores to be used interchangeably, although "close" is not specifically defined.

Dorans and Holland (2000) suggest evaluating the degree to which a linkage function varies across different populations to assess the degree to which it is appropriate to use the linkage between two scores. The measures of population invariance in the linkage function presented by Dorans and Holland (2000) only depend on observed variables — no latent variable model is involved. The criteria discussed by Dorans and Holland (2000) depend on the extent to which $F_Y(y \mid z)$ and $F_X(x \mid z)$ differ for different populations z.

Using the concordant individual scores $t(Y)$ in place of the scores X can result in different inferences being made when the deviations from first and second order equity are not small. The greater the deviations from first and second order equity are from zero the greater the chance of inferences using $t(Y)$ being different from inferences using X. It is likely, due to the fact that Y and X are nonequable, that there will be significant deviations from first or second order equity for $t(Y)$ (no matter what transformation t is used). It will in general not be appropriate to use individual concordant scores $t(Y)$ in place of individual scores X when X and Y are nonequable.

The only condition under which it may be appropriate to use individual concordant scores is when there is specific evidence that the inferences to be made using the concordant scores are likely to be valid. If the concordant score $t(Y)$ will be used to make inferences in place of the actual score X, evidence of the appropriateness of using $t(Y)$ in place of X would involve the use of data from examinees who have taken both Y and X to show the inferences using $t(Y)$ are highly similar to the inferences that are made using X.

For example, one use of a SAT to ACT concordance would be to concord SAT cutoff scores to be used with ACT scores. Since the concordance of SAT to ACT is computed using an equipercentile function, the percentile rank of a SAT cutoff y_c should be approximately the same as the percentile rank of the corresponding ACT score x_c, where

$t(y_c) = x_c$. It is not the case though that the *same* examinees would be classified above the cutoff using a SAT score as would be using an ACT score. Therefore, if a college wished to transform an SAT cutoff to be used with ACT scores, evidence would be needed that a high proportion of examinees would be classified the same way whether their SAT or ACT scores were used. A way of evaluating the appropriateness of transforming cutoffs using concordances is to use probabilities of consistent classification at a cutoff using concordant and actual scores. These probabilities of consistent classification are denoted *consistency rates.*

For a concordant score $t(Y)$ the conditional consistency rate for observed score x given latent variable value $\boldsymbol{\theta}$ is

$$CR(x, \boldsymbol{\theta}) = A(x, \boldsymbol{\theta}) + C(x, \boldsymbol{\theta}) = 1 - [B(x, \boldsymbol{\theta}) + D(x, \boldsymbol{\theta})] ,$$

where

$$\begin{aligned} A(x, \boldsymbol{\theta}) &= \Pr(X \geq x, t(Y) \geq x \mid \boldsymbol{\theta}) \\ B(x, \boldsymbol{\theta}) &= \Pr(X \geq x, t(Y) < x \mid \boldsymbol{\theta}) \\ C(x, \boldsymbol{\theta}) &= \Pr(X < x, t(Y) < x \mid \boldsymbol{\theta}) \\ D(x, \boldsymbol{\theta}) &= \Pr(X < x, t(Y) \geq x \mid \boldsymbol{\theta}) . \end{aligned}$$

These four probability regions are depicted graphically in Figure 2. The deviation from equity given by Equation 1 can be written as

$$\begin{aligned} DE(x, \boldsymbol{\theta}) &= B(x, \boldsymbol{\theta}) + C(x, \boldsymbol{\theta}) - [C(x, \boldsymbol{\theta}) + D(x, \boldsymbol{\theta})] \\ &= B(x, \boldsymbol{\theta}) - D(x, \boldsymbol{\theta}) . \end{aligned}$$

Therefore, the conditional consistency rate can be written in terms of the deviation from equity as

$$CR(x, \boldsymbol{\theta}) = 1 - [DE(x, \boldsymbol{\theta}) + 2D(x, \boldsymbol{\theta})] .$$

The consistency rate is the expected value of conditional consistency rate over $\boldsymbol{\Theta}$:

$$E_{\boldsymbol{\theta}}[CR(x, \boldsymbol{\theta})] = E_{\boldsymbol{\theta}}[A(x, \boldsymbol{\theta})] + E_{\boldsymbol{\theta}}[C(x, \boldsymbol{\theta})] = 1 - E_{\boldsymbol{\theta}}[B(x, \boldsymbol{\theta})] - E_{\boldsymbol{\theta}}[D(x, \boldsymbol{\theta})] .$$

Computation of consistency rates requires a group of examinees who have taken both tests. For a value of score X (say x_0), let y_0 be a value of score Y such that $t(y_0) = x_0$. The consistency rate corresponding to x_0 is the proportion of examinees whose score on X is greater than or equal to x_0 and score on Y is greater than or equal to y_0, or whose

score on X is below x_0 and score on Y is below y_0. This is the consistency of classification using cutpoint x_0 on score X, and cutpoint y_0 on score Y given by the concordance.

Tables 1 and 2 list the consistency rates for some of the tests for which the equipercentile concordance has been done. For each of these examples, the half-at-or-below definition of percentile rank was used to compute the concordance (see Pommerich, Hanson, Harris, & Sconing, 2000, for a discussion of various ways of defining percentile rank). Table 1 gives the consistency rates for the ACT Mathematics test and four ASSET Mathematics tests and the ASSET Writing Skills test (the ASSET tests are designed to be used for college placement), based on an ASSET to ACT concordance where concordant ACT score points were established for each ASSET score point (consistency rates for five concordances are presented in Table 1). The values listed are the consistency rates for the concordant ACT Mathematics score given at the left and an ASSET score. For example, both and 42 and 43 on the ASSET numerical skills test concord to an ACT mathematics score of 18. The entry of 0.78 in the ASSET numerical skills column of Table 1 corresponding to an ACT mathematics score of 18 means that 78% of the examinees taking both ACT and ASSET had both an ASSET score greater than or equal to 42 and an ACT score greater than or equal to 18, or had both an ASSET score less than 42 and an ACT score less than 18.

The blank entries in Table 1 correspond to scores where there was no concordant equivalent (i.e., no concordant ASSET score was equal to that ACT score). The correlations between each ASSET score and the ACT Mathematics score are given in the last row. Appropriately, the lower the correlation, the lower the minimum consistency rate across ACT score points. Also note that, as expected, the Writing Skills test generally gives the lowest consistency rates and has the lowest minimum consistency rate.

Table 2 lists the consistency rates for two separate forms of the ACT composite. The data consisted of students who had taken the ACT Assessment on two national test dates, the first time in April and again the following October. This table gives an idea of the largest consistency rates that can be expected.

Consistency rates need to be interpreted with caution due to factors that may result in the consistency as reported in the tables being higher than it actually is for some purposes. First, a consistency table is, in a strict sense, only appropriate for the population from which the sample used to construct the table was taken. The consistency results may not hold for a group that is quite different from the population for which the table was

constructed. Second, the consistency rates are typically computed using the same data used to compute the concordances, which may result in the consistencies being overstated. A better assessment of consistency would be to compute the consistency table using a cross-validation sample, different from the sample used to construct the concordance.

Consider the case in which a consistency table is computed for an ACT to SAT concordance that shows a high level of consistency. The inference that is validated by the consistency table is that of a concordant ACT score being used in place of a SAT score to determine whether an examinee's SAT score is greater than or equal to a particular cutoff. Thus, an examinee's ACT score can be validly translated to be either greater than or equal to the SAT cutoff or less than the SAT cutoff. If the concordance were being used for a college admission process which included a cutoff on SAT scores it would be appropriate to use the information for ACT-tested applicants that their SAT scores obtained from the concordance were either greater than or equal to or less than the SAT cutoff. It would not be appropriate to use the ACT-tested applicant's concordant SAT scores, just whether those concordant SAT scores exceed the cutoff or not.

Another example of how inferences to be made from a concordant score could be validated is given by the following hypothetical situation. Suppose two forms of a test are used as a pretest and posttest to assess change. We want to assess change for examinees who have taken the pretest, but have taken a different test that is not parallel to the pretest in place of the posttest. This situation was simulated using data on 84,260 examinees who took the ACT Assessment both on the October 1998 and April 1998 test dates. The gain in Reading scale scores for individuals between April and October will be used as the statistic of interest. The effect of using an October Science Reasoning scale score in place of an October Reading scale score to assess Reading scale score gain from April to October will be examined. Both the Science Reasoning and Reading tests are passage–based. Scale scores on the Science Reasoning and Reading tests have a fairly high correlation (0.75), which is comparable to the correlation for some tests for which concordances are computed.

A random sample of 1500 examinees from the total group of 84,260 examinees was used to compute an equipercentile concordance from the October Science Reasoning scale score to the October Reading scale score (scales scores on both the Reading and Science Reasoning tests range from 1 to 36). A concordance table was created that gave concordant Reading scale scores corresponding to each Science Reasoning scale score. In addition, this estimation sample of 1500 was used to compute a linear regression to predict October

1998 Reading scores from October 1998 Science Reasoning scores. The performance of these concordances for the purposes of using October Science Reasoning scores in place of October Reading scores for assessing gain in Reading scores was evaluated using a cross-validation sample of 1500 (the examinees in the estimation and cross-validation samples were mutually exclusive).

Figure 1 presents plots of the gain in Reading scores as computed using the April and October Reading scores (actual gains), and as computed using the April Reading score and the concordant October Science Reasoning score (concordant gains). The top plot in Figure 1 gives results using concordant Science Reasoning scores computed using an equipercentile function. The bottom plot in Figure 1 gives results using concordant Science Reasoning scores computed using regression. The number of observations at each point in Figure 1 is roughly indicated by the size of the plotted symbols. The three symbol sizes indicate 1-5 observations, 6-10 observations, and greater than 10 observations (with larger symbols corresponding to more observations). The line in each of the plots represents an identity line, on which the points would fall if the concordant and actual gains were the same.

The spread of concordant gains at each level of actual gain is rather large. For instance, at an actual gain of zero the concordant gains range from about -10 to 10, which is about as wide as the range of values observed for the actual gains across examinees. Thus, an examinee with an actual gain of zero may have a concordant gain that is about as low as the lowest actual gain or about as high as the highest actual gain. The standard error of measurement for the Reading test is about 2.5 scale score points, so the standard error of the difference in two independent administrations is about 3.5. The spread of gains of concordant scores at each level of actual gain is large relative to the spread that would be expected based on measurement error in the April and October Reading scores. For the concordant scores based on both an equipercentile function and regression, the concordant gains are shifted toward zero from the actual gains at the extremes. So individuals with high (positive) actual gains will tend to have lower concordant gains, and individuals with low (negative) actual gains will tend to have higher concordant gains. The results in Figure 1 suggest that it would not be appropriate to use the individual concordant gains in place of the actual gains.

One procedure that has been used to minimize the possibility of incorrect inferences being made when using concordances with individual scores is to report a range of scores

rather than a single concordant score. For example, in 1989 an Enhanced ACT was introduced and a concordance table was developed which gave concordant Enhanced ACT scores corresponding to original ACT scores (an example in the next section contains more details concerning the Enhanced ACT and original ACT). Besides a concordant Enhanced ACT score at each original ACT score, a range of Enhanced ACT scores was also provided. It was recommended that the Enhanced ACT score ranges be used when an individual examinee wished to know approximately how they would score on the Enhanced ACT given their score on the original ACT. The use of score ranges may help to minimize inappropriate inferences being made if it is likely the concordance results will be used with individual scores.

An example in which a range of concordant scores could be used in applying concordance results to individual scores for placement purposes involves a two-stage decision rule based on a "decision zone." ASSET is a test designed for college placement decisions. Suppose a school has a cutoff score of 42 on the ASSET Writing Skills test for admission into the standard English course, and a concordance is available that associates a range of ACT English scores with each ASSET Writing Skills score. A decision zone strategy allows use of the ACT English score to place students who have ACT scores, minimizing the number of ACT-tested students who would also need to take ASSET. Suppose that an ASSET Writing Skills score of 42 corresponds to a range of ACT English scores from 16 to 18. An example of a decision zone rule would be to place a student with an ACT English score of 15 or below into the remedial course, and place a student with an ACT English score of 19 or above into the standard course. Students with an ACT English score in the decision zone of 16, 17, or 18 would take the ASSET Writing Skills test, and would be placed based on their score on that test. This method requires more testing than just using the concordance to obtain concordant ASSET scores from ACT English scores, but it leads to fewer incorrect placements due to differences between the tests. In the above example, only those students who scored 16–18 on the ACT English test would retest with ASSET and be placed using their ASSET scores, those who scored less than 16 or greater than 18 would be placed using their ACT English scores.

Evaluation and Appropriate Uses of Concordant Score Distributions

Equation 8 shows that deviation from equity of score distributions is the average of the deviation from equity of individual scores. The deviation from equity of score distributions will likely be closer to zero than the deviation from equity of at least some individual

scores. Hence, using concordant score distributions is potentially less problematic than using concordant scores for individual examinees. The appropriateness of using concordant score distributions for a particular population depends on the extent to which the deviation in Equation 6 is zero for that population.

A use of concordance as applied to a score distribution is the computing of norms. An example of this is the concordance developed to allow norms to be computed for the Enhanced ACT Assessment. In 1989, ACT introduced what was then called the Enhanced ACT Assessment, which was the first major revision of the ACT Assessment since its introduction some 30 years earlier. Because of the changes in content, particularly the addition of the Reading and Science Reasoning tests and the retirement of the Social Studies Reading and Natural Sciences Reading tests, scores on the original ACT (administered before October 1989) and the Enhanced ACT (administered in October 1989 and later) could not be considered interchangeable. However, there were substantial similarities in the two assessments: in format, in overall difficulty, in their close ties to the high school curriculum, and in their purposes. The Enhanced ACT and original ACT were similar enough that it seemed reasonable to attempt to maintain/establish a linkage between scores on the two tests, but the tests were dissimilar enough that the linkage would need to be treated as a concordance.

It was determined that the primary purposes of such a concordance would be generating norms. One set of ACT Assessment norms are generated for the graduating class in a given year, using the most recent set of scores for all graduating seniors who were administered the ACT Assessment. As the Enhanced ACT was introduced in October of 1989, it was possible some seniors graduating in May of 1990 would have most recently taken the ACT as juniors (prior to October, 1989). In order to include those students in the graduating class norms, their original ACT scores would need to be linked to the Enhanced ACT Assessment score scale. Concordance tables were developed linking the original ACT to the Enhanced ACT using a nationally representative sample of examinees from a study conducted in the fall of 1988. An equipercentile function was used to compute the concordance table giving the Enhanced ACT scores corresponding to the original ACT scores. The concordance was applied to the distribution of original ACT scores for members of the 1990 graduating class who had only original ACT scores (i.e., did not test as seniors) for the purpose of including them in the graduating class norms. The appropriateness of this procedure depends on how well Equation 6 is satisfied for the population

of ACT-tested students in the 1990 graduating class who did not test in their senior year, when using the concordance developed with the fall 1988 nationally representative sample.

The appropriateness of using concordant score distributions for a particular population depends on the extent to which Equation 6 holds for that population. Evidence that Equation 6 approximately holds in a number of populations similar to the one in question could be used as evidence that it is appropriate to use the concordant distribution. On the other hand, caution would be needed in applying the concordances computed from a sample pooled across institutions to individual institutions, especially when the group of students at an institution differ greatly from the full sample. See Pommerich, Hanson, Harris, and Sconing (2000) for a discussion of this topic.

Discussion

Three levels of comparability between two scores Y and X along a continuum of score comparability were described: closely equable (the scores are measures of the same thing and are from forms that are designed to be parallel), weakly equable (the scores are measures of the same thing but are from forms that are not designed to be parallel), and nonequable (the scores do not measure the same thing). A linking function $t(Y)$ transforms Y in an attempt to make the deviation from equity of $t(Y)$ and X smaller than the deviation from equity of X and Y, although is unlikely that the deviation from equity of $t(Y)$ and X will be *substantially* smaller than the deviation from equity of Y and X. For this reason it is argued that linkage types are most clearly distinguished by the level of score comparability of the scores to be linked (e.g., closely equable, weakly equable, nonequable). The most commonly used categorization of linkage types (Mislevy, 1992; Linn, 1993) mixes level of score comparability with the procedure used to compute the linking function (three of the linkage types correspond to different procedures for linking nonequable scores).

In this paper the term "concordance" is used to refer to a linking function computed for nonequable scores. The deviation from equity for nonequable scores was presented which defines the extent to which when computing a concordance of Y to X it is appropriate to consider $t(Y)$ interchangeable with X for individuals. The deviation from first and second order equity is unlikely to be small when X and Y are nonequable, no matter what linking function is computed. This is a widely accepted conclusion (e.g., Angoff, 1964; Lindquist, 1964). For example, it is probably in general not appropriate when using the ACT to SAT concordance table (or the SAT to ACT concordance table) to treat concordant

scores of individuals as interchangeable with actual scores without considering random and systematic errors in the concordance.

It is only appropriate to use individual concordant scores in situations where empirical evidence exists that the specific inferences to be made using concordant scores $t(Y)$ will not be too different from the inferences made using X. Some examples were presented of the types of empirical investigations that might be carried out to verify that concordant scores result in valid inferences being made. Separate evidence of the validity of using individual concordant scores needs to be obtained for each inference for which the concordant scores are to be used even if such validity evidence exists for the actual score. This is in contrast to equating in which a high level of equity exists, and equated scores are deemed valid to use for a wide range of inferences for which validity evidence for the test in question exists (an equated score is used interchangeably with the score it is being equated to).

The deviation from equity for concordant score distributions will be smaller than the deviation from equity of individual concordant scores. The use of concordant score distributions is likely to be less problematic than the use of concordant individual scores. Still, it is important to have evidence that the deviation from equity of score distributions is approximately zero in any population for which the concordance is used that differs greatly from the population in which the equipercentile concordance function was computed.

References

Angoff, W. H. (1964). Technical problems of obtaining equivalent scores on tests. *Journal of Educational Measurement, 1*, 11-13.

Bloxom, B., Pashley, P., Nicewander, W. A., & Yan D. (1995). Linking to a large-scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics, 20*, 1-26.

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University, 73*, 24-33.

Dorans, N., & Holland, P. W. (2000). *Population invariance and the equatability of tests: Basic theory and the linear case.* Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, April).

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press.

Hanson, B. A. (1993). *Equipercentile equating with equal interval scores.* [Available at http://www.b-a-h.com/papers/note9301.html].

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions.* ETS Technical Report 89-84. Princeton, NJ: Educational Testing Service.

Houston, W., & Sawyer, R. (1991). Relating scores on the enhanced ACT Assessment and SAT test batteries. *College and University, 66*, 195-200.

Kolen, M. J., & Brennan, R. L. (1995) *Test equating methods and practices.* New York: Springer.

Lindquist, E. F. (1964). Equating scores on non-parallel tests. *Journal of Educational Measurement, 1*, 5-9.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83-102.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Marco, G. L., & Abdel-Fattah, A. A. (1991). Developing concordance tables for scores on the enhanced ACT Assessment and the SAT. *College and University, 66*, 187-194.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, N.J.: Educational Testing Service.

Morris, C. N. (1982). On the foundations of test equating. In P. W. Holland and D. B. Rubin (Eds.), *Test equating* (pp. 169-191). New York: Academic.

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2000). *Issues in creating and reporting concordance results based on equipercentile methods.* ACT Research Report 2000-1. Iowa City, IA: ACT, Inc.

Williams, V. S. L., Rosa K. R., McLeod, L. D., Thissen, D., & Sanford, E. E. (1998). Projecting to the NAEP scale: Results from the North Carolina end-of-grade testing program. *Journal of Educational Measurement, 35*, 277-296.

Table 1
Consistency Rates for ACT Mathematics and ASSET Tests

Concordant ACT Mathematics score	ASSET test				
	Numerical Skills	Elementary Algebra	Intermediate Algebra	College Algebra	Writing Skills
11	0.99				0.99
12	0.96	0.97	0.98		0.97
13	0.91	0.91	0.95	0.98	0.91
14	0.84	0.84	0.90	0.95	0.85
15	0.77	0.76	0.83	0.92	0.76
16	0.75	0.72	0.77	0.89	0.70
17	0.75	0.72	0.73	0.83	0.68
18	0.78	0.75	0.73	0.80	0.69
19	0.83	0.79	0.75	0.77	0.73
20	0.87	0.84	0.78	0.77	
21	0.89	0.87	0.81	0.77	0.79
22	0.91	0.91	0.86	0.78	0.84
23	0.93	0.93	0.89	0.79	
24	0.94	0.95	0.91	0.82	0.90
25	0.96	0.96	0.94	0.85	0.92
26	0.97	0.98	0.96	0.87	0.95
27	0.98	0.99	0.97	0.90	
28			0.98	0.93	0.97
29	0.99	0.99	0.99	0.95	0.98
30				0.96	
31	1.00		1.00	0.98	
32					
33				0.99	
Correlations	0.70	0.62	0.66	0.72	0.52

Table 2
Consistency Rates for ACT Composite on Different Test Dates

ACT Composite	Consistency Rate
10	1.0
11	1.0
12	1.0
13	0.99
14	0.98
15	0.97
16	0.95
17	0.93
18	0.91
19	0.90
20	0.89
21	0.89
22	0.89
23	0.89
24	0.89
25	0.90
26	0.91
27	0.92
28	0.94
29	0.95
30	0.96
31	0.98
32	0.99
33	1.00
34	1.00
35	1.00
36	1.00

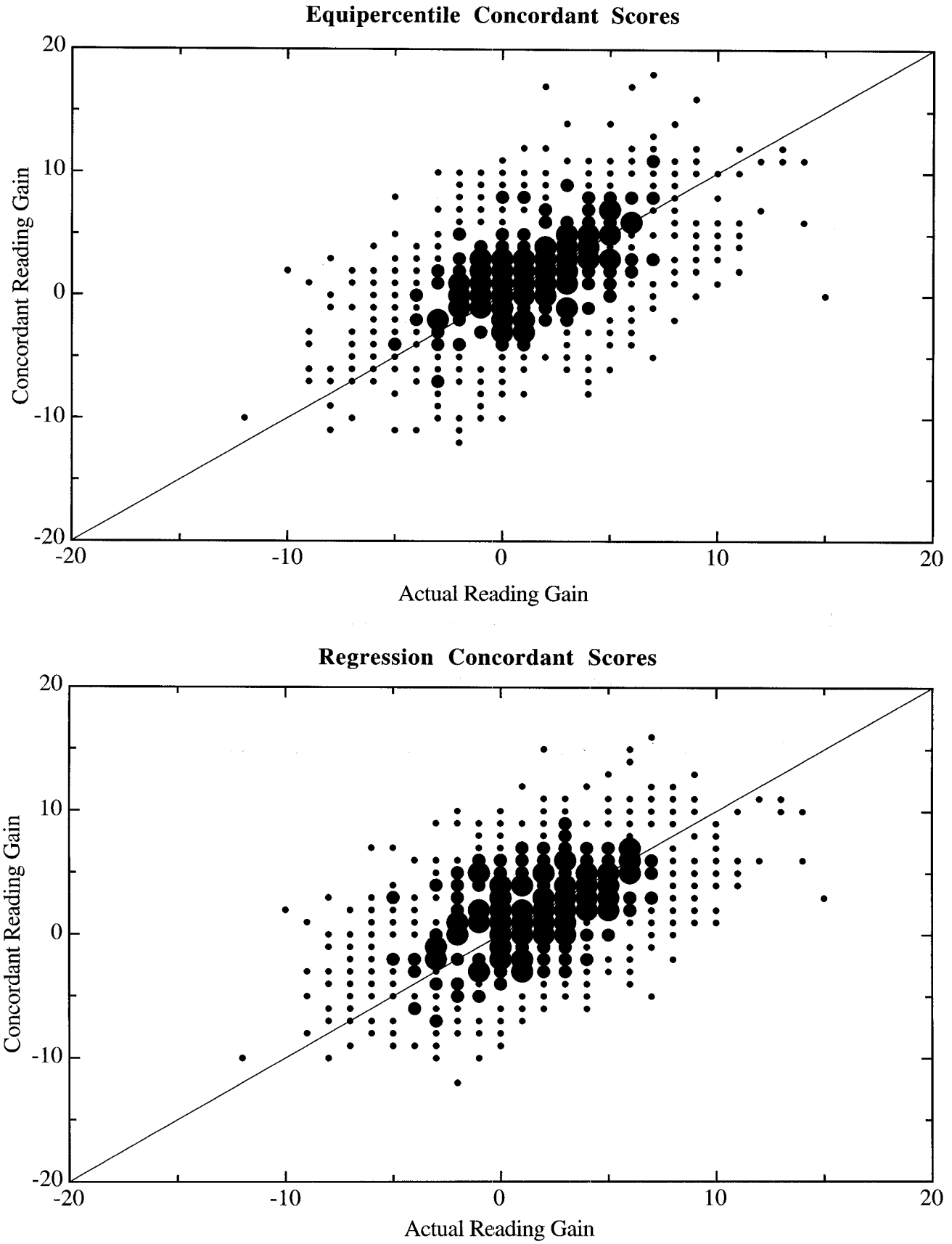

Figure 1. Actual Reading Gain Versus Concordant Reading Gain Using October Science Scores.

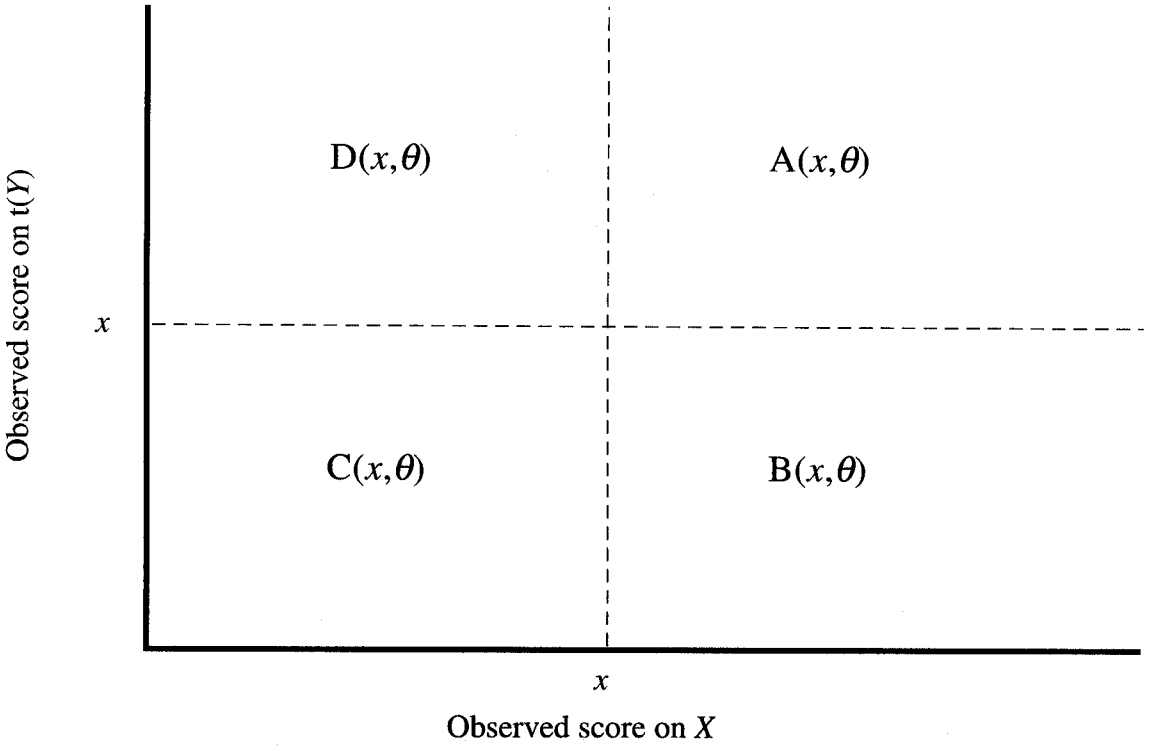

Figure 2. Probability Regions for Conditional Consistency Rate.