The College Board

Research Notes

Office of Research and Development RN-11, December 2000

# Distinctions Among Classes of Linkages

Neil J. Dorans

Score users often wonder how different tests relate to each other. Some users are cautious and hesitate to make comparisons across tests. Others presume that all tests can be linked in a manner that leads to simple comparisons and valid inferences. To some, “different tests” means two versions of the same test that are built to a clearly specified blueprint. To others, “different tests” means measures of the same construct (e.g., math) built to different specifications (e.g., those used by ACT for the ACT test, and those used by Educational Testing Service [ETS], for the College Board’s SAT® I test). To others, “different tests” is used to make a distinction between a test of math knowledge and a test of science reasoning.

Users of test scores often like to use scores interchangeably. Sometimes they presume that the scores are completely exchangeable. To ensure that scores are compared in the proper way, a better understanding of the continuum that ranges from strict exchangeability of scores to no association between scores is needed.

Several authors—for example, Angoff (1971), Linn (1993), and Mislevy (1992)—have discussed distinctions among different types of score linkages. The current paper presents a conceptual framework for linkages between scores and score scales that distinguishes among three kinds of linkages, namely, equating, scaling, and prediction.

Construct similarity plays an important role in determining the degree of linkage that can be achieved. This paper also maintains that statistical indices in conjunction with rational considerations are needed to determine whether the highest level of linkage attainable between scores from two “tests” is the conceptual and statistical exchangeability sought by equating, the distributional similarity of scaling, or the association attained by prediction.

KEYWORDS:

linking
uncertainty reduction
equating
prediction

Relationships among the different scales of the ACT and SAT I, two nationally known college admission tests, are described in the context of the conceptual framework developed herein. Users want to know how scores on the ACT and the SAT I are related. Dorans, Lyu, Pommerich, and Houston (1997) presented linkages between SAT I and both the ACT Sum and the ACT Composite. Data from that study are used to evaluate the appropriateness of concordances and predictions among various scores on these two prominent tests. Sums of scores, composites of scores, and individual scores are examined. Different types of linkages between different sets of scores from these two admission tests are amenable to different kinds of interpretations.

## CLASSES OF TEST SCORE LINKAGE

Three classes of linkage are delineated in this paper: equating, scaling, and prediction.

### Equating

The goal of equating (Holland & Rubin, 1982; Kolen & Brennan, 1995) is to produce scores that are fully exchangeable. A score is exchangeable with another score if it is a measure of the same thing, say length, and expressed in the same metric, say inches, as another score. The two scores may have been obtained via two versions of the same measuring instrument. A simple example is the length of a piece of string. Most foot-long rulers are gradated in inches and centimeters. If we measure a string in both metrics, we can easily convert the string’s length “scores” into the same metric, either centimeters, inches, feet, or meters. The point is that length is the construct being measured and that meters, inches, feet, or miles are all fully equatable; i.e., they can be placed on the same metric. Scores on tests of developed abilities and skills can be equated too, provided they are constructed to the same set of specifications, and a proper data collection design can be used to establish the equating relationship (Angoff, 1971). Imperfect reliability prevents test scores from achieving the equatability associated with virtually infallible measures, such as length.

### Scaling

A second type of linkage between two scales is scaling. Typically, the data collection designs and the statistical techniques used to establish a scaling relationship are also used to establish an equating relationship. The crucial distinction is that two scales that have been placed on a common metric are considered equated only if they measure the same thing. For example, different editions of the SAT I are placed on the same scale with the intent of producing exchangeable scores. An examinee should be able to take any edition of the SAT I and get the same reported scores on the 200 to 800 scale within the precision (reliability) of the test. The same can be said for ACT scores. SAT I scores and ACT scores, however, are not exchangeable. They measure different constructs. When SAT I V+M (a sum) and ACT Sum (or ACT Composite) are scaled to each other, as they recently were by Dorans, Lyu, Pommerich, and Houston (1997), concordance tables are produced. Because the correlation between the ACT Sum and SAT I V+M was so high (.92), scaling was used in the Dorans et al. (1997) study to establish the linkages between these sum scores. This means, for example, that the score on ACT Sum that corresponded to the same percentile in some group as a score on SAT I V+M was denoted as corresponding or concordant. This does not mean, however, that scores on ACT Sum and SAT I V+M are exchangeable. Likewise, a scaling of SAT I Verbal to SAT I Math does not yield exchangeable scores.

One distinguishing characteristic of scaling (and equating) is that the relationship between the two scores is invertable. That means that if a 125 on ACT Sum corresponds to a 1400 on SAT I V+M, then a 1400 on SAT I V+M corresponds to a 125 on ACT Sum. This statistical equivalence does not mean that a 125 and a 1400 can be used interchangeably as measures of the same construct. Instead, they can be thought of as occupying the same location in a rank ordering of scores in some group of people.

### Prediction

The third type of linkage to be discussed is prediction. It is the least restrictive and least demanding type of linkage. Whereas equating strives to achieve fully exchangeable scores and scaling matches distributions of scores, prediction is merely concerned with doing the best job possible to predict one set of scores from another. The goal is to minimize the imprecision in the predictions of one score from one or more scores. A classic example of a prediction model is the estimation of grade-point average from earlier grades and high school scores. Unlike scaling and equating relationships, prediction relationships are not symmetric; i.e., the function that converts scores on test A to scores on test B is not the multiplicative inverse of the function that converts scores on test B to scores on test A.

## WHICH TYPE OF LINKAGE?

How do we know the degree to which we can achieve exchangeability, concordance, or prediction? There are three factors that provide us with an answer in any given situation.

First, we must perform a logical evaluation of the similarity of the processes that produced the scores to see if the constructs measured are similar. Second, we need to assess the strength of the empirical relationship between the scores that we wish to link. Typically this relationship is measured by the correlation coefficient. Third, we must assess the degree to which a linkage relationship is invariant across subpopulations.

To achieve the exchangeability of equating, the tests must be measuring the same construct, the correlation between the two tests must be high, and the linkage relationship must hold across important subgroups.

## An Example: Differences and Similarities in ACT and SAT® I Content Specifications

Different editions of the SAT I are constructed to be similar in content and difficulty by experienced assessment professionals who use a clearly specified blueprint to guide them. These tests are administered to students seeking admission in colleges and universities. The rigor of the assembly process and the motivation of the students taking the tests combine to produce scores that can be equated. ACT uses its professional assembly process, and administers its tests to comparably motivated students to produce scores that are also equatable. The two processes, though different in some ways, yield distributions of sum scores that are highly correlated, and can be related via concordance tables.

The process used to produce grades differs markedly from those used to produce test scores. In contrast to test scores, which are obtained from carefully constructed tests administered under standardized conditions in a brief period of time, grades are a cumulative record obtained under varied non-standard circumstances. Prediction is the best that one can expect under these circumstances. The relatively low correlations between grades and test scores, compared to those obtained among test scores, attest to the dissimilarity of these processes.

## A Measure of Uncertainty Reduction

To support scaling, the correlation must be high. If the correlation is too low then prediction is the only option.

McNemar (1969) describes a vintage statistical index (Kelley, 1919) called the coefficient of alienation that is a measure of statistical uncertainty that remains after inclusion of information from the predictor variable. This index involves the correlation coefficient, *r:*

$$\text{coefficient of alienation (coa)} = \surd(1\text{-}r^2).$$

We can define the reduction of uncertainty as:

$$\text{reduction of uncertainty} = 1 - \text{coa} = 1 - \surd(1\text{-}r^2).$$

Note that when $r$ equals zero the coefficient of alienation equals one, which means that there is a zero reduction in uncertainty about scores on the measure to be predicted. For example, if the information in the predictor variable (say a randomly picked lottery number) has no relationship with variation in scores on the variable to be predicted (the change in wealth expected to occur as a result of the draw of the winning number), then the predictor does nothing to reduce my uncertainty about performance on the variable to be predicted (winning the lottery). In contrast, a 100 percent reduction of uncertainty, represented by a zero coefficient of alienation, is achieved when $r = 1$.

A 50 percent reduction is halfway between 100 percent reduction ($r$=1) and 0 percent reduction ($r$=0). A correlation coefficient of at least .866 is needed to reduce the uncertainty, as measured in score units, of knowing a person's score by at least 50 percent. If a predictor cannot reduce uncertainty by at least 50 percent, it is unlikely that it can serve as a valid surrogate for the score you want to predict.

Figure 1 plots reduction of uncertainty (*y*-axis) as a function of the correlation (*x*-axis). Substantial uncertainty reduction requires large correlations as indicated by the slow-to-rise slope for correlations below .80 and the steep slope above .80.

The selection of any cutpoint is arbitrary, but it may or may not be capricious. What does a 50 percent reduction in uncertainty mean in

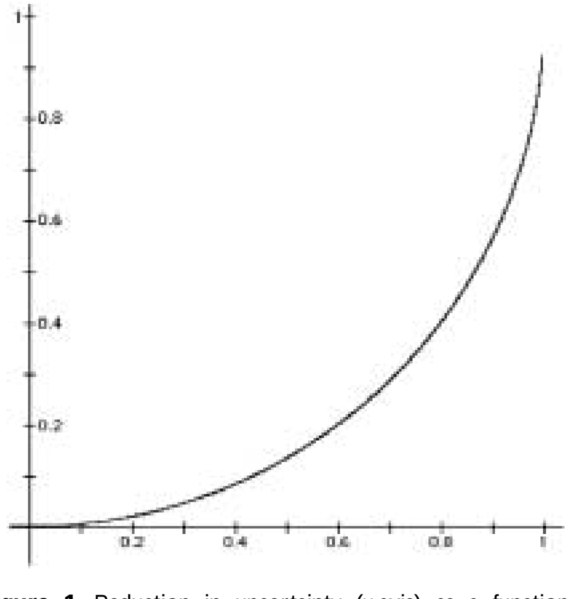

**Figure 1.** Reduction in uncertainty (*y*-axis) as a function of correlation (*x*-axis).

concrete terms? Suppose we were asked to predict a person's height, and all we knew was that he or she was an adult. With no other information, our best guess would be the average height of an adult, and the standard deviation of height among adults would represent an uncertainty measure of one. If we knew she was female, our estimate would shift downward, and our uncertainty measure would get smaller. As we added more and more information about her, such as age, weight, height of parents, etc., our uncertainty would continue to reduce.

To better appreciate the reduction of uncertainty, consider the notion of test reliability. If we know nothing about a man other than that he took a test along with a group of other men, we could use the average male score to estimate his performance on the test. In this case, the standard deviation of that group would represent an uncertainty of one. If we knew his true ability (something we can't know), we could use this true score as an estimate of his observed score on any test. A true score could be used with a test of reliability .75 to reduce the uncertainty of observed score performance by 50 percent, which means a test with a reliability of .75 has a standard error of measurement equal to half the original standard deviation.

The squared correlation between true and observed score is one definition of reliability. The correlation between two parallel test forms is an equivalent definition of reliability. Note in Figure 1, however, that a correlation between parallel forms of .87 is needed to reduce uncertainty in predicting one observed score from another observed score by 50 percent. In other words, a true score can reduce uncertainty as well on a test with reliability of .75 as an observed score can for two parallel tests of reliability .87.

For the SAT I and ACT in the population studied by Dorans et al. (1997), if an examinee presents an ACT Composite score, it reduces uncertainty about his or her SAT I V+M score by 60 percent, because the correlation between ACT Composite and SAT I V+M is .92. In other words, the range of plausible SAT I V+M scores is reduced by 60 percent once we have knowledge of an examinee's ACT Composite score. The logical evaluation needs to be verified with the empirical data. Reductions in uncertainty that fall short of 50 percent may be indicative of scores that are neither equivalent nor concordable. To illustrate the role of the correlation in assessing concordability, we will examine the SAT I and ACT composites and component scores.

Before proceeding, it is necessary to address anticipated criticism of using the correlation coefficient. The correlation coefficient has its limitations. For one, it does not describe non-linear relations well. For the purposes of this paper, we will assume that the distributions of the two scores have either been matched (Holland and Thayer, 2000) or are similar enough in shape that a linear relationship is adequate for prediction purposes. Another criticism is that correlation coefficients can be easily attenuated. For example, suppose we are only interested in distinguishing among SAT I Math scores at 750 or above. ACT Math scores or scores from any other measure, including those from another edition of SAT I Math, would not be of much use because the range restriction on the score we are interested in predicting is so severe that virtually all potential predictors have very limited validity. The attenuated correlation reflects these practical limitations. The fact that it suggests that two versions of SAT I Math are not correlated enough to warrant exchangeability is a troublesome, but accurate, description of what is achievable in the highly restricted subpopulation of data under study.

## Population Invariance

Equatability is an important aspect of equity. The goal of equating is to ensure that scores on one test can be used interchangeably with scores from another test. Assessment of equatability provides a framework from which to evaluate whether changes are minor or substantial.

Lord (1980) specifies four prerequisites for equating:

1. The two tests must measure the **same construct**;
2. The equating must achieve **equity**; i.e., for individuals of a given proficiency, the conditional distributions of scores on each test must be equal;
3. The equating transformation should be **symmetric**; i.e., the equating of $Y$ to $X$ should be the inverse of the equating of $X$ to $Y$; and

4. The equating transformation should be **invariant** across subpopulations of the population on which it is derived.

Population invariance is an important requirement because equating transformations are one-to-one relationships between scores that should be unique and identical across subpopulations from the population. If population invariance is not achievable, it may be due to the fact that the tests are not measures of the same construct. For example, a scaling of SAT I Verbal scores to SAT I Math scores would not qualify as an equating because different relationships would occur for males and females. Checking the equivalence of equating relationships across subpopulations is a sure way of assessing the population invariance requirement. The absence or presence of population invariance distinguishes a scaling (absence) from an equating (presence of invariance).

In the current study, we use an indirect index, namely the standardized difference between male and female means, of what would be obtained if we actually scaled the two tests in the two gender populations. Dorans and Holland (2000) demonstrate that if the scaling of the tests were actually parallel-linear in both populations and the standardized differences between males and females were equal on both tests, then population invariance would hold because linear scaling is performed in terms of these standardized differences, or differences in standard deviation units. To the extent that the scaling relationship is not well approximated by a linear scaling, this simple standardized difference in means will be a misleading approximation.

## RESULTS

### Content Comparison

The SAT I yields two scores: a Verbal score based on 78 questions administered in 75 minutes, and a Math score based on 60 questions administered in 75 minutes. The ACT yields six scores: an English score based on 75 questions administered in 45 minutes, a Math score based on 60 questions administered in 60 minutes, a Reading score based on 40 questions administered in 35 minutes, and a Science Reasoning score based on 40 questions given in 35 minutes.

At this very general level of description, the ACT Math score and the SAT I Math score appear similar in name and number of questions. In contrast, the SAT I and the other three ACT scores all appear different. Further evaluation of the content specifications of the tests that produce these six scores confirm these apparent similarities and differences. More than $^5/_6$ths of the SAT I Math content comes from three primary domains: arithmetic, algebra, and geometry. Less than $^1/_6$th of the mathematics items are drawn from other areas of mathematics, such as trigonometry. For ACT Math, about $^{11}/_{12}$ths of the items come from algebra, geometry, and pre-algebra. Trigonometry items make up the balance of the test. To the extent that “arithmetic” and “pre-algebra” overlap, the specifications between SAT I Math and ACT Math are quite similar. This high level of content linkage suggests strong statistical concordance.

The SAT I Verbal measures verbal reasoning via critical reading (about half the test), analogical reasoning questions, and sentence completions.

ACT English measures the elements of effective writing. About half the test is dedicated to usage/mechanics of the English language, which is assessed via punctuation, grammar and usage, and sentence structure. The remainder of the test assesses rhetorical skills, i.e., strategy, organization, and style. The content of ACT English is similar to the writing test that used to be administered with the old SAT, the Test of Standard Written English. It measures something more akin to the SAT II: Writing Test than it does the SAT I Verbal test.

ACT Reading, in fact, is more aligned from a content perspective with the SAT I Verbal than is ACT English. The questions in this test come from four domains: prose fiction, social sciences, humanities, and natural sciences. It appears as if this test measures half of what the SAT I Verbal test measures, namely the reading portion of reasoning.

ACT Science Reasoning measures science knowledge via three formats: data representation, research summaries, and conflicting viewpoints. It does not appear to be aligned with any other test.

Table 1 contains a condensed comparison of the various ACT and SAT I component scores. SAT I Math and ACT Math are contained within

**TABLE 1**
**CONTENT COMPARISON OF SAT I AND ACT SCORES**

| | | | | | | |
|---|---|---|---|---|---|---|
| SAT I Verbal | Critical Reasoning (36-44) questions | Analogies/ Sentence Completions (34-42) questions | | | | |
| SAT I Math | Arithmetic Reasoning (18-19) questions | **Algebraic** Reasoning (17) questions | | **Geometric** Reasoning (16-17) questions | | Miscellaneous Reasoning (7-9) questions |
| ACT Math | Pre-Algebra (14) questions | Elementary **Algebra** (10) questions | Intermediate **Algebra** (9) questions | Coordinate **Geometry** (9) questions | Plane **Geometry** (9) questions | Trigonometry (4) questions |
| ACT English | Usage/ Mechanics (40) questions | Rhetorical Skills (35) questions | | | | |
| ACT Reading | Prose Fiction (10) questions | Humanities (10) questions | | Social Sciences (10) questions | Natural Sciences (10) questions | |
| ACT Science Reasoning | | | | Research Summaries (18) questions | Conflicting Viewpoints (7) questions | Data Representation (15) questions |

Source: College Board's *Handbook for the SAT Program* (1996-97) and ACT's *Test Preparation Reference Manual.*

bold lines to highlight their similarity. SAT I Verbal is set apart from the four ACT scores to emphasize its dissimilar content. This comparison of the content of the two SAT I tests and the four ACT tests suggests that a solid linkage should be found between the mathematical portions of ACT and SAT I, and does not suggest any other likely concordances, with the possible exception of SAT I Verbal and ACT Reading.

## Empirical Relationships Among ACT and SAT I Scores

Dorans, Lyu, Pommerich, and Houston (1997) describe the processes used to screen data and select the concordance sample used for the analyses that linked the composite scores for ACT and SAT I. Their description covers data collection, data screening, matching of data files, the effects of time between testings on test performance relationships, and other factors.

The relationship between ACT and SAT I scores was evaluated for students taking both tests between October 1994 and December 1996, and within 217 days of each other. The scaling sample consisted of student records from 2 states and 14 universities. The samples for states and for institutions were mutually exclusive, so that a student was represented in either the state sample or the institution sample, but not both. The sample used for this study was not a random sample of all students who took both examinations. The total number of student scores used in the analyses was 103,525 students.

The scaling procedure used by Dorans et al. (1997) was the equipercentile method. A single group design was used in which students took both forms to be scaled. As the name implies, the equipercentile method sets equal the scores that have the same percentile ranks in the sample. For example, the 90th percentile in the ACT Sum score distribution is set equal to the 90th percentile in the SAT I V+M score distribution. See Dorans et al. (1997) for a discussion of technical issues associated with using equipercentile equating with these data.

**Uncertainty Reduction**

In addition to scaling the tests, measures of reduction in uncertainty, based on the Pearson product moment correlation coefficient, were computed.

Dorans, Lyu, Pommerich, and Houston (1997) reported a correlation between SAT I V+M and ACT Sum (Composite) of .92 in a sample of 103,525 students who took both the SAT I and ACT. The magnitude of this correlation justified the reporting of concordances back and forth between the SAT I and ACT composites, which are reported in Dorans et al. (1997), and Dorans (1999).

Correlations among individual SAT I and ACT scores were also computed. As suggested by the content analysis above, the highest correlation of any ACT score with any SAT I score was the .89 between ACT Math and SAT I Math. SAT I Verbal correlated .83 with ACT Reading and .83 with ACT English. The equivalence of these statistical relations mirrors the ambiguity about how the SAT I Verbal relates to these two ACT scores that we observed in the content analysis. The fact that ACT Reading and ACT English correlate .81 with each other is further evidence that SAT I Verbal, ACT English, and ACT Reading are distinct measures.

Correlations in the low .80s are considered high, especially in the context of predicting grades from test scores, where the unreliability of the grade point average and its scaling problems attenuate the correlation coefficient. But in establishing linkages between test scores, correlations in the low .80s are too low to merit concordance tables, and unacceptable if the goal is to establish exchangeability of scores.

In fact, we suggested earlier that correlations below .866 reduce uncertainty by less than 50 percent; hence, their scores are not sufficiently concordable. The correlations observed for the ACT Science Reasoning test indicate the need to draw the line somewhere near the mid .80s. This test correlates .76 or .75 with each of the other SAT I and ACT scores. Few would argue that this ACT Science Reasoning measure is a measure of SAT I Math, SAT I Verbal, ACT Reading, ACT English, and ACT Math all at the same time. Nor would many argue that SAT I V+M is a measure of ACT Science Reasoning because it correlates .82 with it. Likewise, few would argue that SAT I V+M is a measure of ACT English because it correlates .87 with it. Most, however, would agree that these ample correlations would yield solid predictions of performance on these other tests. Prediction, yes. Concordance, no. Exchangeability, definitely not.

Expectations based on content considerations are verified when the reduction of uncertainty index is used. These results are summarized in Figure 2.

Figure 2 summarizes the types of linkages that work best for ACT and SAT I scores. There are eight circles, one for each of the three SAT I scores and five ACT scores. At the top of the figure are two circles for composite or sum scores, SAT I V+M and ACT Composite/ACT Sum. (ACT

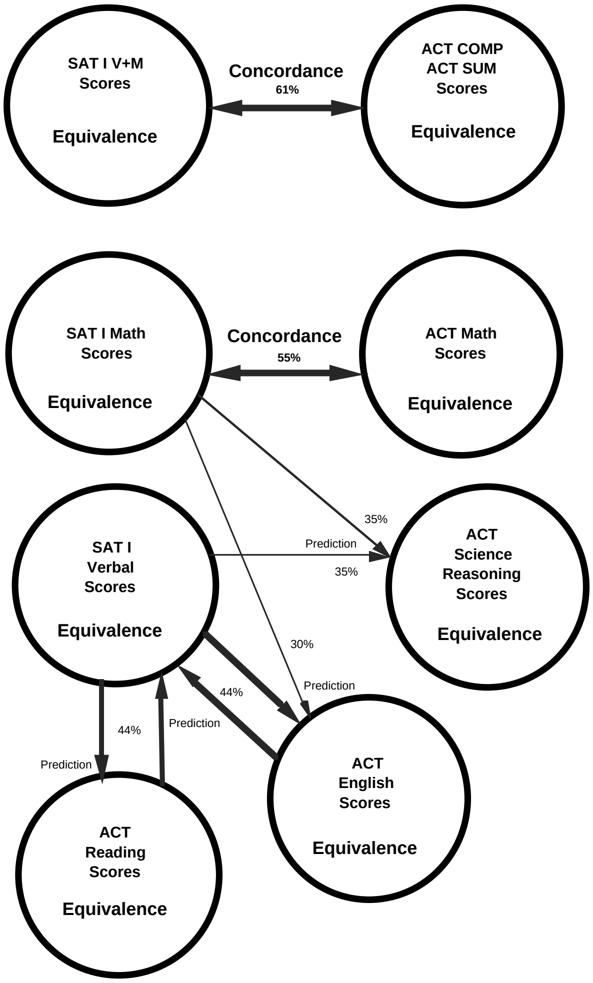

**Figure 2.** ACT/SAT I linkages.

Sum and ACT Composite are included in the same circle because the Composite is simply the Sum divided by 4 and rounded to the nearest integer.) Just below the composite/sum circles are two circles for the SAT I and ACT Math scores. The remaining four circles are for SAT I Verbal and the three other ACT scores: Reading, English, and Science Reasoning.

The word equivalence appears within each of the eight enclosed circles to denote that scores within these circles are designed to be exchangeable with each other. For example, all ACT Math scores come from test editions built to the same specifications and are equated in an effort to achieve exchangeability across different editions of the ACT. Likewise, all SAT I Verbal scores come from test editions built to the same specifications and are equated so that they can be used interchangeably regardless of which edition they came from. Equivalence is the strongest form of linkage that scores can have, and bounded circles denote a closed system of score equating that is designed to yield exchangeable scores.

The sum/composite circles for ACT and SAT I are connected by a line that is bi-directional and labeled concordance. These bi-directional lines connecting two distinct circles denote a strong statistical relationship between sets of scores drawn from tests built to different sets of specifications. ACT Math and SAT I Math scores are also represented by a concordance relationship. Different groups could have different concordant relationships. In contrast, equating relationships are the same across different groups. In short, concordant scores cannot be used interchangeably in the way equivalent scores can be. ACT Math scores and SAT I Math scores, though highly related, should not be used interchangeably.

A second type of arrow appears in the figure. This unidirectional arrow denotes a prediction relationship. For example, SAT I Math in conjunction with SAT I Verbal can be used in a prediction equation to predict ACT English scores. A predicted ACT English score is the score that people who have a certain combination of SAT I scores are expected to get on ACT English. Actually, there is an expected range of ACT English scores associated with each pair of Math and Verbal SAT I scores.

Prediction is also the preferred form of linkage for ACT Reading, as indicated by the unidirectional arrow that goes from SAT I Verbal to ACT Reading. Prediction is used in this case instead of concordance because the correlation of .83 is not high enough for an equipercentile concordant relationship to give users a reasonable estimate of the expected ACT Reading score given their SAT I scores. Note that there is no arrow connecting SAT I Math to ACT Reading because SAT I Math does not add much to the predictability of ACT Reading beyond that already associated with SAT I Verbal.

A prediction model can also be used for ACT Science Reasoning. Unidirectional arrows lead from the SAT I Math and SAT I Verbal circles to the ACT Science Reasoning circle. Use of a prediction model for ACT Science Reasoning seems reasonable because concordance between SAT I Math and ACT Science Reasoning or between SAT I Verbal and ACT Science Reasoning does not make sense given the definition of concordance. There are few if any requests for concordances between ACT Science Reasoning and SAT I scores.

In contrast, there are many requests for concordances between the two math scores. There are also requests for concordances between ACT English and SAT I Verbal. Perhaps these requests stem from the misconception that SAT I Verbal is an English test or that ACT English is a Verbal Reasoning test. Actually, as we learned from our examination of the content specifications, ACT English is a writing skills test. Perhaps it is highly related enough to SAT II: Writing to warrant a concordance relationship. The data used in this study indicate that it is no more related to SAT I Verbal than ACT Reading is, which is not surprising because reading items comprise about half of SAT I Verbal. Therefore, no concordant relationship exists between SAT I Verbal and ACT English or ACT Reading. Likewise, it makes little sense to establish a concordance between ACT English and ACT Reading.

### Population Invariance

Population invariance separates equating relationships from scaling relationships. There is a simple way to determine if invariance is unlikely or likely to hold across subpopulations: compute standardized differences between important subpopulations. Gender differences are an obvious choice.

To compute standardized differences of each test between means for males and means for females, the female mean was subtracted from the male mean, and the result divided by the total group standard deviation. Since about 57 percent of the sample was female, the female group had a slightly larger impact on the total group standard deviations.The results are illustrated by Figure 3.

First, the obvious will be discussed. Any scaling of SAT I Math to SAT I Verbal will differ across male and female subpopulations. Likewise any scaling of ACT Math (or ACT Science Reasoning) to either ACT English or ACT Reading would not be the same for males and females.

Second, the SAT I Math, ACT Math, and the Science Reasoning tests have similar standardized gender differences. This convergence suggests that the two math tests might yield nearly exchangeable scores, a conclusion that is not contradicted by either the content or uncertainty reduction analyses. This finding can be viewed as convergent validation for both measures.

Third, SAT I Verbal and ACT Reading have more similar standardized gender differences than either has with ACT English, the only score with a negative standardized difference. The dissimilarity of ACT English and SAT I Verbal is further confirmed with these data.

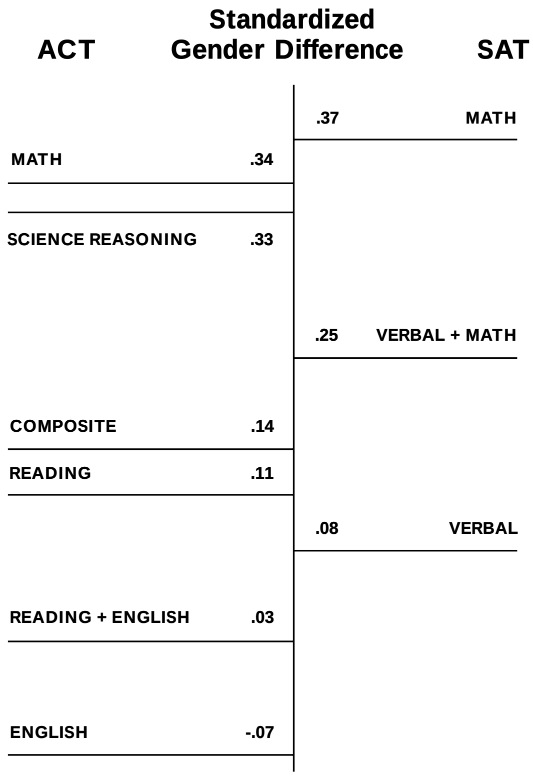

**Figure 3.** Standardized mean differences (male-female) for candidates taking both ACT and SAT I.

Finally, the standardized difference of a composite or sum score is simply a function of differences on the scores that define it. ACT Composite has a lower standardized difference because of ACT English. If an ACT composite is defined as the sum of ACT Math and ACT Reading, its standardized difference would be close to that observed for SAT I V+M, and the degree of concordability between SAT I V+M and this particular ACT composite might be higher than it currently is.

In sum, the standardized mean difference results buttress the concordance relationship between SAT I Math and ACT Math, diminish the relationship between ACT Composite and SAT V+M, and confirm that prediction is the most appropriate linkage for the remaining test scores. These results also confirm that SAT I Verbal and ACT Reading are more aligned to each other than either is to ACT English.

## SUMMARY

Distinctions were made between three classes of statistical linkage: equivalence, concordance, and prediction. These distinctions were based on rational content considerations and empirical statistical relationships. A large database involving SAT I and ACT scores was used to determine which type of linkage was best suited for different scores and composite scores.

Equating is used routinely within each testing program.

Earlier research had produced concordance tables between the ACT Composite/Sum and the SAT I sum (Dorans et al., 1997), and between SAT I Math and ACT Math (Dorans, 1999; Maxey, 1998). The current research provides a content-based and empirical justification for these concordances.

Applying the same rationale to the SAT I Verbal and ACT Reading, ACT English and ACT Science Reasoning scores led to the conclusion that these scores are not concordable from either a content or statistical rationale. Prediction is the more appropriate form of linkage for these scores.

SAT I Verbal is best predicted by ACT English and ACT Reading scores. ACT Reading is best predicted by SAT I Verbal (the Math score adds little to this prediction). Both ACT English and ACT Science Reasoning are best predicted by a combination of SAT I Verbal and SAT I Math.

We assessed three factors to arrive at these conclusions. First, we performed a logical evaluation of the similarity of the processes that produced the scores to see if the constructs measured were similar. Second, we assessed the strength of the empirical relationship between the scores that we wished to link. We used reduction in uncertainty to assess this factor. Third, we assessed the degree to which a linkage relationship is invariant across subpopulations by examining standardized difference in male and female means. This process can be applied to other tests in other situations.

*The author is Neil J. Dorans, Principal Measurement Statistician, Educational Testing Service.*

## ACKNOWLEDGMENTS

This paper evolved over the last several years. It is based on earlier work on ACT/SAT concordances done with C. Felicia Lyu and Mary Pommerich that was suggested by Gary Marco, James Maxey, and Richard Sawyer. It started as a response to concerns raised by Howard Everson about linking SAT I scores with ACT subscores. Along the way, comments by Wayne Camara, Michael Kolen, Gerald Melican, Amy Schmidt, Denny Way, and Cathy Wendler helped tighten the presentation. Related work with Paul Holland helped shape the final product.

## REFERENCES

Angoff, W. A. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.), pp 508-600. Washington, DC: American Council on Education. (Reprinted as W. A. Angoff [1984]. *Scales, norms and equivalent scores.* Princeton, NJ: Educational Testing Service, 1984.)

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores.* (College Board Report # 99-1; ETS RR 99-2) New York: The College Board.

Dorans, N.J. & Holland, P.W. (2000) Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4).

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(2), 24-34.

Holland, P. W., & Rubin, D. B. (1982). *Test equating.* New York: Academic Press.

Holland, P.W., & Thayer, D.L. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 2*, 133-183.

Kelley, T.L. (1919). Principles underlying the classification of men. *Journal of Applied Psychology, 3*, 50-47.

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

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.

Maxey, J. (1998, April). *Concordant scores on the ACT and the SAT I: Subject area scores & conclusions.* Paper presented at the annual meeting of the American Association of Collegiate Registrars and Admissions Officers, Chicago, IL.

McNemar, Q. (1969). *Psychological Statistics* (4th ed). New York: John Wiley & Sons.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

**For more information or additional copies of this report, please write to: Office of Research,
The College Board, 45 Columbus Avenue, New York, NY 10023-6992, or contact us by e-mail at:
research@collegeboard.org, or visit our Web site at: www.collegeboard.com.**

12/00
989747