Linking Assessments: Concept and History

Michael J. Kolen, University of Iowa

In this article, the history of linking is summarized, and current linking frameworks that have been proposed are considered. Key publications discussed include Flanagan (1951), Angoff (1971), Linn (1993), Mislevy (1992), and Feuer, Holland, Green, Bertenthal, and Hemphill (1999). The article further focuses on the concordance situation for linking and discusses the concept of concordance and future research that is needed. *Index terms: calibration, comparability, concordance, equating, linking assessments, scaling*

Scores on educational assessments are linked statistically so that scores from one assessment can be expressed in the units of another assessment or so that scores from both assessments can be expressed on a common score scale. In this article, the term *linking* is used broadly, referring to a relationship between scores on two tests. Often, the scores on the assessments to be linked are measures of the same constructs, such as when scores on alternate forms of the Mathematics test of the ACT Assessment are equated. Sometimes, however, scores on tests that measure somewhat different constructs are linked for practical reasons.

For example, colleges often want to admit students who have taken either the SAT I or the ACT Assessment. Consider an oversimplified situation in which a college requires a minimum score of 1200 on the SAT I Verbal plus SAT I Mathematics for admission. If this college accepts the ACT Composite score in place of SAT I scores, the college needs to establish the ACT Composite score that is comparable to the minimum SAT I score of 1200. Comparable scores are found using a linking process.

The term *concordance* is used in this article to refer to linking scores on assessments that measure similar (but not identical) constructs and in which scores on any of the linked measures are used to make a particular decision. The preceding example in which students are allowed to submit either ACT or SAT I scores for college admissions purposes is a typical concordance situation.

Open questions regarding concordance include the following: What is meant when two tests are said to measure similar constructs? How is the similarity of constructs best assessed and quantified? How similar do the constructs need to be before concordance is feasible? What are the consequences of conducting a concordance study when the constructs differ? What are some of the practical problems that arise in designing a concordance study and in analyzing the results? Some of these questions are addressed in this article and subsequent articles in this special issue.

In this article, the history of linking is summarized, and current linking frameworks that have been proposed are considered to provide a context for concordance. The fit of concordance into linking frameworks is discussed, as is future research that is needed for concordance.

Applied Psychological Measurement, Vol. 28 No. 4, July 2004, 219–226
DOI: 10.1177/0146621604265030

Flanagan (1951)

Flanagan (1951) provided an in-depth discussion of linking and the construction of score scales. He used the term *comparability* to refer to scores on tests that were scaled so that they had similar score distributions in a certain population of examinees. He stated, "Scores . . . on two or more tests may be said to be *comparable* for a certain population if they show identical distributions . . . for that population" (p. 699). He discussed comparability of scores on multiple test forms as follows:

> Different forms of the same test must yield comparable scores if they are to be completely interchangeable in use. Such interchangeable forms of a single test that yield comparable scores are called *equivalent* forms . . . . Equivalent forms of . . . tests are not only closely equated in the process of construction, but they are also accompanied by standard score scales which bring about more precise comparability. (p. 699)

Flanagan (1951) stated that "forms which are truly interchangeable in that they measure the same functions with equal accuracy and are reported in comparable scores will be designated as 'equivalent forms' " (p. 748). He stressed that for scores on multiple test forms to be able to be used interchangeably, the forms must be constructed to be similar, and a score scale must be used to express the scores.

Situations in which comparable scores could be constructed for tests that were constructed to measure different constructs were also discussed. He stated that "if results from several tests of different types are to be compared, as in preparing a profile of results from a battery of tests taken by an individual, it is necessary that the scores from the various tests have some type of meaningful comparability" (Flanagan, 1951, p. 699). One example that he used was to construct scales for different tests in a battery (e.g., a reading and a mathematics test) to have the same distributional characteristics for a particular population. He considered this type of comparability useful for constructing profiles to assess relative strengths and weaknesses.

For both types of comparability, Flanagan (1951) indicated that the relationship between scores was population specific. He stated, "Comparability which would hold for all types of groups—that is, *general* comparability between different tests, or even between various forms of a particular test—is strictly and logically impossible" (p. 748). There is some suggestion by Flanagan that comparable scores for multiple forms of a given test could be expected to be less population specific than comparable scores for different tests.

In discussing the issue of linking scores from tests that differ in difficulty (e.g., levels of an achievement test battery appropriate for different grades), Flanagan (1951) again used the term *comparability* to refer to this process. Regarding the relationship between comparability and test reliability, he further indicated that to establish comparable scores, the distributions of true scores should be the same for the measures. In this regard, he stated that "if the reliability of measurement is the same for two tests for the population involved, then similar results will be obtained if the distributions of obtained values are compared" (p. 747). If tests are not similar in reliability, he suggested that the true score distributions should be scaled to be comparable.

In discussing the use of regression methods to relate scores on different tests, Flanagan pointed out that regression procedures "do not give *comparable* scores" (Flanagan, 1951, p. 751). He illustrated the lack of comparability by demonstrating, through an example, that regression is not symmetric, and he discussed the practical effects of the lack of symmetry on score interpretation.

In summary, Flanagan (1951) distinguished the process of linking scores on test forms constructed to measure the same construct from linking scores on tests that measure different constructs. He discussed the process of linking scores on tests that measure the same construct but differ

in difficulty, incorporated test reliability into his conception of score comparability, distinguished symmetric linking functions from nonsymmetric regression functions, and considered all linking functions to be population dependent. Although he considered many of the situations and issues that are of interest today, as the subsequent discussion suggests, most of the terminology that is in current use was developed after Flanagan's (1951) effort.

Angoff (1971)

Angoff (1971) summarized and organized the work on linking and scaling that had been conducted up to the time of writing. He used the term *equating* to refer to linking scores on multiple forms of a test in which the alternate forms were built to the same specifications to be as similar as possible to one another. Unlike Flanagan (1951), Angoff stated that equating relationships should be population independent.

Angoff (1971) used the term *calibration* to refer to linking scores on tests that measure the same construct but differ in difficulty or reliability. Thus, linking scores on tests from an elementary school achievement battery that are designed for different grade levels, which is often referred to as *vertical scaling*, would be referred to as *calibration*, using his terminology.

He used the term *comparability* to refer to linking scores on tests that measure different constructs. Although Flanagan (1951) had used the term *comparability* to refer to any relationship between scores in which the score distributions were scaled to be similar, Angoff (1971) restricted the use of the term.

Similar to situations discussed by Flanagan (1951), Angoff (1971) discussed comparability when scores from different tests in a test battery are scaled to have a common distribution to assess an examinee's strengths and weaknesses relative to a norm group. He also described a more complicated but related situation in which scores on the College Board Achievement tests (now known as the SAT II tests) were scaled. Examinees chose which achievement tests to take, so the ability level of the group of examinees who took the achievement tests differed from one test to another. Each examinee who took an achievement test also took the SAT, so the relationship between the SAT and the achievement tests could be estimated for the examinees. Angoff described a process in which a hypothetical group of examinees was formed for which the SAT scores for this group were 500 and the standard deviation 100. Scores on each of the achievement tests was scaled to have a mean of 500 and a standard deviation of 100 for this hypothetical group. Strong statistical assumptions were required to conduct the scaling. However, the process was intended to model the scaling that would have occurred had all examinees in the hypothetical group been administered each of the College Board Achievement tests as part of a test battery.

Angoff (1971) also considered a situation in which scores on tests from different test publishers were linked, as would be the case with the ACT Assessment and SAT I, as described in the introduction of this article. He indicated that these linkings were population dependent. He stressed that

> this nonuniqueness of comparable scores derives from the fact that the measures in question are measures of different function; there is no unique single conversion table that can be derived for tests of different function, and there is no unique single conversion that is applicable to all types of groups. (p. 595)

He indicated that the usefulness of tables of comparable scores depends on the answers to two questions: "How similar are the tests for which the comparable scores are to be developed?" and "How appropriate is the group on which the table of comparable scores is based when one considers the person or the group for whom the table is to be used?" (Angoff, 1971, p. 597).

Mislevy (1992) and Linn (1993)

Mislevy (1992) and Linn (1993) developed a conceptual framework for linking that includes four types of statistical linking: equating, calibration, statistical moderation, and projection (prediction). They also discussed social moderation, which is primarily a judgmental (not statistical) process. This framework is referred to as the Mislevy/Linn framework in this article. In their framework, consistent with Angoff (1971), *equating* refers to linking scores on alternate forms of an assessment that are built to common content and statistical specifications. Also consistent with Angoff (1971), the term *calibration* is used when scores are linked on tests that are intended to measure the same construct but with different levels of reliability or different levels of difficulty.

Projection and *statistical moderation* are used in the Mislevy/Linn framework to refer to linking when the tests measure different constructs. In their framework, projection involves predicting scores on one test from scores on another using regression methodology. To conduct a projection study, the same examinees typically are administered both of the tests that are linked.

The term *statistical moderation*, which was also used by Keeves (1988), derives from a process in which scores on each test are linked to a third (moderator) variable. The process used to scale scores on the College Board Achievement tests described earlier is an example of statistical moderation. Typically, for statistical moderation, one group of examinees takes one test and the moderator variable, and a different (often nonequivalent) group of examinees takes the other test and the moderator variable. For example, the College Board French and Spanish Achievement tests could be scaled using SAT Verbal as a moderator variable.

Mislevy (1992), but not Linn (1993), also referred to a simpler moderation situation in which two tests are given to the same group, or randomly equivalent groups, of examinees. One example of this type of statistical moderation would be to link ACT and SAT I scores using observed score-equating methods based on a group of examinees who took both tests. The process of linking ACT and SAT I scores described in the introduction of this article is an example of this type of statistical moderation.

Mislevy (1992) and Linn (1993) stressed that statistical moderation and projection results are group dependent. For statistical moderation, the transformation that is developed is symmetric; for projection, the transformation is not symmetric.

Scores on any measures can be linked using statistical moderation or projection from the Mislevy/Linn framework, as long as appropriate data are collected. There is nothing in their conception that can be used to distinguish a situation in which scores on tests to be linked measure similar constructs from a situation in which scores on tests to be linked measure very different constructs.

Uncommon Measures (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999)

The National Research Council studied the feasibility of linking scores on achievement tests administered by the states to scores on the National Assessment of Educational Progress (NAEP). The final report (Feuer et al., 1999) broadly reviewed the problem of linking scores on tests and adapted and further developed the framework developed earlier by Mislevy (1992) and Linn (1993).

The NAEP is intended for group-level score interpretation; no scores are reported to individuals. Examinees take only a portion of the assessment. The content of an NAEP assessment is viewed as the content over all portions of the test. In addition, the NAEP often contains a large proportion of constructed response questions. Under many circumstances, the broad content coverage and large number of constructed response items would be impractical in a test administered to individuals.

To help consider content differences between tests to be linked, Feuer et al. (1999) focused on three stages in test development that they referred to as follows:

- *framework definition*—a delineation of the scope and extent (e.g., specific content areas, skills, etc.) of the domain to be represented in the assessment;
- *test specification or blueprint*—specific mix of content areas and item formats, number of tasks/items, scoring rules, and so on; and
- *item selection*—items are selected to represent the test specifications as faithfully as possible.

Using this conceptualization of test development, three types of linking were discussed. *Equating* refers to linking when the same framework and the same test specifications are used to construct the test forms whose scores are to be linked. *Calibration* refers to linking when tests are built to the same framework but with different test specifications. Vertical scaling is one example of calibration. When the framework differs, then *moderation* or *projection* is used. The terms *equating*, *calibration*, *moderation*, and *projection*, as used by Feuer et al. (1999), are similar to their use by Mislevy (1992) and Linn (1993). Feuer et al., however, attempted to tie their use directly to the test development process.

Feuer et al. (1999) also took into account other various special features of the NAEP. The NAEP is administered by a single contractor, leading to highly standardized and tightly controlled administration. A much greater level of standardization is used with the NAEP than with typical state-level testing programs. NAEP scores are reported as group-level statistics based on estimated distributions of proficiencies that contain no measurement error. Because individuals do not receive scores, NAEP performance has no direct consequences for an individual.

Consideration of these special characteristics of the NAEP led Feuer et al. (1999) to conclude that various factors affect the validity of links. These factors include the following for the tests whose scores are to be linked: (a) the similarity of content, difficulty, and item formats; (b) the comparability of measurement error associated with the scores; (c) the test administration conditions; (d) the uses being made of the tests and the consequences of test use, including the stakes associated with test results; and (e) the accuracy and stability of the links, including the stability over subgroups of examinees.

Dorans (2000, 2004 [This Issue])

Dorans (2000, 2004) makes a distinction between the linking of scores on tests that measure different constructs from those that measure similar constructs. Dorans (2004) refers to the linking of scores on tests that measure similar constructs as *concordance*, in which the statistical methods lead to similar score distributions for the measures. Dorans (2004) suggests that concordance be used when tests are measuring similar constructs, their content is judged to be similar, scores are highly correlated, and linking relationships differ very little from one group of examinees to another. He suggests that relationships meeting these criteria can be expected to provide comparable scores on two measures useful for certain decisions and appropriate for a wide range of examinee groups. He concludes that regression methods should be used to link scores on measures that cannot be related using concordance procedures.

Kolen and Brennan (in press)

As indicated earlier in this article, Feuer et al. (1999) suggested that many aspects of linking situations are important when evaluating linking. Kolen and Brennan (in press) proposed a scheme for

categorizing many of these aspects in terms of what he refers to as features of the linking situation. The four features that were proposed are as follows:

- *Inferences:* To what extent are scores for the two tests used to draw similar inferences?
- *Constructs:* To what extent do the two tests measure the same constructs?
- *Populations:* To what extent are the two tests designed to be used with the same populations?
- *Measurement conditions:* To what extent do the tests share common measurement conditions, including, for example, test length, test format, administration conditions, and so on?

Kolen and Brennan (in press) discuss how these features relate to categories of linking in the other schemes. For example, with equating, the alternate forms whose scores are equated are used to make the same inferences and measure the same constructs, are designed to be used with the same populations, and are administered under the same conditions. For a concordance between scores on ACT and SAT I Mathematics tests, the tests are designed to make similar inferences and measure similar constructs, are designed to be used with similar populations, and are administered under similar conditions. In linking state tests to the NAEP, there are dissimilarities on all of the features.

Summary of Linking Frameworks

The linking frameworks point to the process of *equating* as being distinct and clearly defined. To be equated, test forms are constructed to the same content and statistical specifications. Following equating, scores can be used interchangeably. Although equating relationships are defined using a population of examinees, equating relationships have been shown to be very similar from one examinee group to another (Kolen, in press).

Calibration refers to the process of placing scores for tests designed to measure the same construct, but in which the tests can differ in difficulty and reliability, on a common score scale. The tests whose scores are to be calibrated are developed using a content framework that helps ensure that the construct being measured is the same from one test to another. One example of calibration is linking scores on a short version of a test to scores on a long version of a test, in which the short version is constructed to have the same proportion of items (with the same statistical properties) in each content category as the longer version. Another example of calibration is to link scores on levels of an elementary school mathematics achievement test. For this type of test, a separate level might be developed at each of Grades 3 through 8, and the calibration process is used to place scores for all of the levels on the same developmental score scale. This particular type of calibration is often referred to as vertical scaling.

When a test battery is developed, often the scale scores for each test are defined to have the same distributional characteristics for a particular population of examinees. The process of constructing such a scale is referred to here as *test battery scaling*. Note that this type of scaling is not directly referred to by Mislevy (1992) and Linn (1993), Feuer et al. (1999), or Dorans (2000, 2004). Yet, the process is a type of linking, is widely used, and can be traced back to Flanagan (1951).

When scores on tests that measure different constructs are linked, the terminology and situation become much more complex. Terminology used by Mislevy (1992) and Linn (1993) refers to a linking conducted using regression methods as projection or prediction and linking conducted using procedures that lead to similar distribution characteristics as *statistical moderation*.

Many different situations exist in which scores on tests that measure different constructs might be linked. Test battery scaling is one of them, which clearly produces useful results in that it allows test users to compare strengths and weaknesses of examinees. A second example is linking scores on state tests to scores on the NAEP. As pointed out by Feuer et al. (1999), however, such a linking is not likely to be useful because there are so many differences between the NAEP and state-level tests

in terms of constructs measured, test uses, administration conditions, and so on. A third example is the linking of scores on the ACT and SAT I, which was described in the introduction of this article. The ACT and SAT I tests are used for similar purposes, have similar consequences, have similar administration conditions, and, at least in mathematics, have similar content. Thus, it seems possible that such a linking between scores on tests from these two batteries might be useful. Of the linking frameworks discussed, only Dorans (2000, 2004) specifically distinguishes linking scores on tests that measure similar constructs from tests that measure different constructs. Dorans (2004) uses the term *concordance* to refer to linking of scores on tests that measure similar constructs.

Feuer et al. (1999) made it clear that a whole host of issues, beyond statistical ones, should be considered when linking scores on tests that differ in the construct being measured. Kolen and Brennan (in press) provided a list of four features of the situation that can be used to evaluate linkings in this broader context.

Research Directions in Concordance

As stated in the introduction of this article, the term *concordance* is used here to refer to linking scores on assessments that measure similar (but not identical) constructs, in which scores on any of the linked measures are used to make a particular decision. One example of a concordance situation is linking scores on ACT and SAT I tests. Another is linking selected scores on achievement tests from different publishers that cover similar content, after a school district decides to change test batteries. A third example is to link scores on a revised test to scores on an original test, in which the revised test was constructed to different specifications from the original test.

The quality and usefulness of a concordance depend on the similarity of the tests and testing situations, broadly defined. Kolen and Brennan's (in press) list of four features provides a starting place for evaluating linkings. The similarity of the two tests should be addressed in terms of the inferences made, the constructs measured, the intended populations, and the conditions of measurement.

Judgmental processes are a component of analyzing these features. For example, the purposes of the tests can be compared when analyzing the inferences to be made. A systematic comparison of test content is an important component in comparing the constructs measured. A careful analysis of measurement conditions should include analyzing the test administration conditions and the stakes associated with test results. In general, there is clearly a need to develop systematic judgmental procedures for analyzing the similarity of tests and testing conditions to assess whether concordance between two tests is likely to be possible and useful.

Statistical analyses are also valuable in evaluating similarity. Estimated correlations between observed scores and between true scores (Lord & Novick, 1968) on the measures can be useful in assessing whether the two tests are measuring the same constructs. Structural equation models (Bollen, 1989) and generalizability theory (Brennan, 2001) analyses can also be helpful in this regard. Test reliability and the pattern of conditional standard errors of measurement for the linked measures also can be compared. Statistical procedures for analyzing the extent that linking is population invariant were proposed by Dorans and Holland (2000). These methods include statistical indices and graphical procedures, which can be helpful in deciding on whether to consider a linking a concordance.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Dorans, N. J. (2000). *Distinctions among classes of linkages* (Research Notes RN-11). New York: The College Board.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227-246.

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), 281-306.

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

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695-763). Washington, DC: American Council on Education.

Keeves, J. (1988). Scaling achievement test scores. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education*. Oxford, UK: Pergamon.

Kolen, M. J. (in press). Population invariance in equating: Concept and history. *Journal of Educational Measurement*.

Kolen, M. J., & Brennan, R. L. (in press). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

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

Author's Address

Address correspondence to Michael J. Kolen, Iowa Testing Programs, College of Education, 224B1 Lindquist Center S, University of Iowa, Iowa City, IA 52242-1529; phone: (319) 335-6429; fax: (319) 335-6038; e-mail: michael-kolen@uiowa.edu.