Linking Scores From Multiple Instruments

Neil J. Dorans
Center for Statistical Theory and Practice
Educational Testing Service

The comparability of measurements made in differing circumstances by different methods and investigators is a fundamental pre-condition for all of science.

Dorans and Holland (2000)

1. Overview

This statement by Dorans and Holland was made in the context of equating educational and psychological tests, but it generalizes to any type of measurement, including health outcomes measurement. To the extent that outcome scores of health assessment instruments are to be used interchangeably, the outcome scores need to be equated or made comparable. If the outcome scores of different health assessment instruments are not equated, inferences based on them could be flawed, and might have serious consequences such as misperceptions about the efficacy of a treatment. It would be nice if outcome scores were comparable as a result of careful instrument design. In reality, that rarely happens without statistical intervention.

This paper addresses key questions associated with the linking of health outcome scores. What is meant by outcome score linking and equating? How does equating differ from other types of linking? What are common data collection designs used to capture data for outcome scores linking? What are some of the standard statistical procedures used to link outcome scores directly? What assumptions do they make? What role does IRT play in linking outcome scores? What assumptions do IRT methods make?

2. What is meant by outcome score linking and equating?

Score equating techniques are those statistical and psychometric methods used to adjust scores obtained on different instruments measuring the same construct so that they are comparable. For example, the SAT I, the well known to many high school students and their parents, is given at several large-scale test administrations each year. A different version or "form" of the SAT is used at each of these large test administrations with the

result that college admissions staff are often in the position of comparing applicants whose SAT I scores come from different versions of this test. Test score equating is designed to eliminate the effect of unintended differences in the relative difficulty of these test versions and makes such across-version comparisons of examinee results meaningful.

Basic references on test equating are Angoff (1971), Holland and Rubin (1982), Petersen, Kolen and Hoover (1989), Kolen and Brennan (1995), and von Davier, Holland and Thayer (2004). The data collections and methods described in these references can be used to link the outcome scores of health assessments, particularly those derived from questionnaires about patient states.

Data collection designs are critical to the score or outcome score linking process. Some techniques require specific data collections. Others can be applied in a variety of data collection settings in which much data are missing, but the price to be paid is strong assumptions about how the missing data would perform if it were available.

Equating and linking methods refer to a collection of techniques that have been developed by creative individuals to solve the score linking problems that have arisen in a wide variety of practical testing circumstances. Most of these techniques divide into two categories: those based on transforming distributions of observed scores and those that make use of "true scores" that are hypothesized to underlie the observed scores. Another important distinction is between linear and equipercentile observed score equating methods.

3. How does equating differ from other types of linking?

In addition to the many techniques for actually doing score linking, there are five "requirements" that are often regarded as basic to all of score equating. These have been translated into health outcomes assessment terms: question replaces item, instrument replaces test, respondent replaces examinee, and outcome score replaces score. The five requirements are:

(a) **The Equal Construct Requirement**: instruments that measure different constructs should not be equated.

(b) **The Equal Reliability Requirement**: instruments that measure the same construct but which differ in reliability should not be equated.

(c) **The Symmetry Requirement**: the linking function for equating outcome scores of instrument *Y* to those of instrument *X* should be the *inverse* of the linking function for equating the outcome scores of *X* to those of *Y*.

(d) **The Equity Requirement**: it ought to be a matter of indifference for a respondent to be tested by either one of two instruments that have been equated.

(e) **Population Invariance Requirement**: the choice of (sub) population used to compute the equating function between the scores of instruments *X* and *Y* should not matter—i.e., the equating function used to link the outcomes of *X* and *Y* should be *population invariant*.

These requirements can be used to distinguish equating from weaker forms of linking, such as concordance, calibration and prediction.

3.1 Concordances. The same calculations used for equating instruments are also used to link the outcome scores that measure the same or similar constructs but according to different specifications. For example, many colleges and universities accept scores on either the ACT or SAT I. Instead of claiming to equate ACT scores to SAT I scores and therefore making the scores on them interchangeable for any purpose, a concordance table or concordance function was produced. This concordance enabled users to better align cut-scores on these two somewhat similar but different tests than they would have been able to using the limited data available to single college or university. Unlike equatings, concordances are more sensitive to the population of examinees whose data are used to estimate the concordance function.

3.2 Calibration. Calibration refers to the process of placing scores on a score scale for tests designed to measure the same construct, but may do so with unequal reliability or unequal difficulty. A content framework is used to ensure that the construct being measured is the same from one instrument to another. A short form of an instrument is less reliable than a longer version and a link between them is an example of a calibration. Another example is vertical linking, where both instruments may be of similar reliability, but of different difficulty, one being targeted for a different population than the other.

3.3 Prediction. Since the 1920's, there has always been a distinction made between the methods related to score equating and those of prediction, i.e., regression. Both approaches may be used to transform scores on one test into the scale of the scores on another test. However, these transformations may be very different and have very different uses. In prediction, the goal is to predict a **Y**-score for an examinee from some other information about that examinee. In prediction there is an inherent asymmetry between what is predicted and what is being used to make the prediction. On the other hand, equating functions do not predict scores on one test from scores on another. Instead, scores that have been equated can be used interchangeably, as if they are from a common test rather than from different tests.

4. What are common data collection designs used to capture data for outcome scores linking?

The role of data collection is crucial to successful instrument linking. It is very important to control for differences in distributions of response propensities when assessing differential instrument difficulty. In test equating or linking, this has always been accomplished through the use of special data collection designs.

All types of outcome score linking are based on definitions, data, and assumptions. Typically data are collected on complete instruments, which means that some group of respondents was administered an intact instrument. Sometimes data on a complete instrument is collected in a systematic piecemeal fashion, and sometimes in a manner that depends on the level of the attribute being assessed. The complete instrument cases will be discussed first.

The single-group design. The single-group design directly controls for differences in response propensities by using the same respondents for both instruments. Special studies, such as those linking the ACT composite to the SAT I V+M score employ this design.

The counterbalanced design. In order to allow for the possibility of order effects in the single group design, the sample is sometimes randomly divided in half and in each subsample the two instruments are taken in different orders—*X* first and then *Y* or *Y* first and then *X*. This design is rarely employed in practice, but has been used often in studies

that examine relationships between tests built to an old set of specifications and tests built to new specifications.

The equivalent groups design. In the equivalent groups design, two equivalent samples are taken from a common population **P**; one is administered instrument *X* and the other instrument *Y*. Obtaining large representative equivalents groups is the key to success with this design. ACT employs this design to place new ACT forms on the 1-36 scale.

These three designs are strong data collection designs because differences in distributions of response propensity are controlled for directly by constructing equivalent groups. The following designs produce weaker data that require more assumptions to produce linking relationships.

Anchor instrument designs. The anchor instrument designs improved upon the flexibility of the equivalent groups design, by allowing the two samples, one from population **P** that is given *X* and one from population **Q** that is given *Y*, to be different or "non-equivalent." However, the two samples can only be different in ways that can be quantified using an anchor instrument, *A*, which is administered to both **P** and **Q**. The statistical role of *A* is to remove bias rather than to increase precision. Because *X* is never observed for examinees in **Q**, and *Y* is never observed for examinees in **P**, some type of assumption is required to "fill in" these missing data. For this reason, there are more methods of equating instruments for anchor instrument designs than there are for the others. All of these methods correspond to assumptions made about the missing data. This design is used to place new SAT I forms on the 200-800 scale.

Incomplete instrument designs. The anchor instrument design collects data for linking instruments that are administered to different groups of respondents. There exist designs for linking instruments that have never been administered to the same set of respondents. These designs range from the highly structured data collections for section pre-equating to designs used for computer adaptive testing. These less structured designs produced weaker data that require stronger assumptions in order to produce links.

5. What are some of the standard statistical procedures used to link outcome measures directly? What assumptions do they make?

Equating or linking methods differ with respect to whether they are linking observed outcome measures or "true" outcome scores, which can be thought of as expected observed scores across many parallel measurements. I focus on observed outcome score linking.

Equating methods differ with respect to what they define as equated outcome scores. In addition to differences between definitions involving observed and true outcome scores, there are differences in the degree to which distributions of outcome scores are matched.

Finally, equating methods differ with respect to how they deal with the missing data. Different assumptions are made by different observed outcome score methods, and the types of assumptions vary with the data collection design.

6. What role does IRT play in linking outcome scores? What assumptions do IRT methods make?

Item response theory can be used to link data collected with any design ranging from the strongest single group design to the weakest design in which only handfuls of items are administered to the same group of people. Its flexibility is a direct consequence of its strong assumptions. IRT produces indirect outcome score linking as opposed to direct linking associated with some observed outcome score linking methods. IRT makes strong assumptions at the level of questions that establishes linkages at the question level. From these question level linkages, indirect linkages among instruments outcome scores can be constructed. Many IRT models share certain assumptions. IRT models assume that their parameterization of the item space and person space produces item parameter invariance across subpopulations of examinees. Many presume that there is only a single person parameter is needed, unidimensionality, and that this person parameter combines with a set of item parameters the describe examinee performance at the item level. As with any model, the assumptions need to be tested.

7. Summary

Equating is defined and contrasted with other forms of outcome score linking. Different data collection designs are described along with methods used with these designs. In general, linkages among outcome scores require either strong data or strong models. Strong data are also preferable to weak data. Weak data requires stronger models than. Anchor instrument equating models make assumptions about instrument outcome scores and anchor instrument outcome scores. IRT models make strong item level assumptions that make it potentially useful in a variety of settings.

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.)

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of equating.* New York: Springer.

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,* 281-306.

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

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

Petersen, N. S., Kolen, M. J., and Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.) *Educational Measurement* (3rd ed., pp. 221-262). New York: Macmillan.