2 A Framework and History for Score Linking

Paul W. Holland[1]

Educational Testing Service and Paul Holland Consulting Corporation

2.1. Introduction

For two tests, a *link* between their scores is a transformation from a score on one to a score on the other. The scores being linked might be raw scores or scaled scores (Angoff, 1971). Linking transformations can be developed in a variety of ways that reflect the similarities and differences between the tests as well as the uses to which the links are to be put. Several frameworks have been suggested for organizing the variety of links that are used in practice. For example, see Flanagan (1951), Angoff (1971), Mislevy (1992), Linn (1993), Feuer, Holland, Green, Bertenthal, and Hemphill (1999), and Dorans (2000, 2004d). In addition, Kolen (2004a) and Kolen and Brennan (2004) reviewed and synthesized several frameworks.

This chapter is concerned with a framework developed in Holland and Dorans (2006) that builds on this prior work. In addition, it gives a brief account of the history of score linking. Along with the next chapter by Kolen, it provides a setting for subsequent chapters in this volume that appear in the part of this volume on equating (Part 2), tests in transition (Part 3), concordance (Part 4), vertical linking (Part 5), and linking scales from group assessments to scales used to report scores on individuals (Part 6).

The term *linking* refers to the general class of transformations between the scores from one test and those of another. Linking methods can then be divided into three basic categories called *predicting*, *scale aligning,* and *equating*. Scale aligning will be shortened to *scaling* when convenient.

[1] The opinions expressed in this chapter are those of the author and not necessarily of Educational Testing Service.

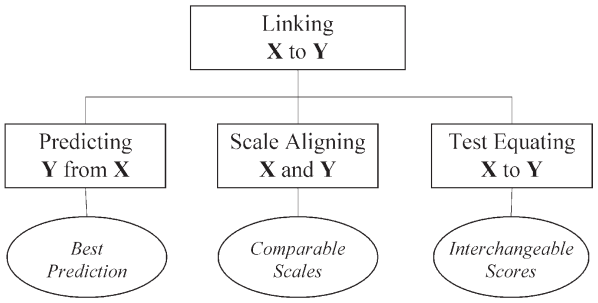

Figure 2.1. The three overall categories of test linking methods and their goals.

Figure 2.1 illustrates the three basic categories of linking and their purposes.

Each of these basic categories contains subcategories that share common objectives and that are distinct from the objectives of the methods in the other categories. It is important to distinguish among these basic categories because they are often seen as similar or identical when in fact they are not. Testing professionals need to understand these differences and the circumstances when one category is more relevant than another and, when necessary, to be able to communicate these distinctions to test users. Figures 2.2, 2.3, and 2.4 illustrate the several subcategories within the basic categories of predicting, scale aligning, and equating.

It is sometimes useful to distinguish between score linkings that are *direct* and those that are *indirect*. A direct link functionally connects the scores on one test directly to those of another. An indirect link connects the scores on two tests through their common connection to a third test or scale. The categories of predicting and equating usually produce direct links, whereas the various subcategories of scale aligning typically produce indirect links. These distinctions are mentioned when appropriate.

2.2. Predicting

Predicting is the oldest form of score linking and it has been confused with the other methods of score linking since the earliest days of psychometrics. By the dawn of the 19th century, Legendre, Gauss, Laplace, and their scientific contemporaries understood how to use least squares methods to fit curves to solve problems in astronomy. By the end of that century,

linear regression methods had been applied to a variety of social and psychological phenomena as well. Notable among these pioneers was Galton, who first observed the effects of regression to the mean (Stigler, 1986). Thus, the use of linear regression methods to predict the scores on one test or measurement from those of another is probably the oldest approach taken for linking scores. A version of predicting, called *projection*, is closely related to certain forms of scaling and equating. Both predicting and projecting are described in this section.

Figure 2.2 illustrates the subcategories within the overall linking category of predicting.

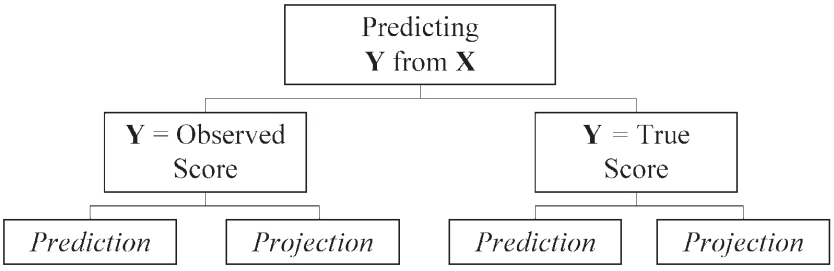

Figure 2.2. The types of linking methods within the overall linking category of *predicting*.

2.2.1. Predicting Observed Scores

The goal of predicting is to predict an examinee's score on one test from some other information about that examinee. This other information might be a score on another test or the scores from several other tests and it might include demographic or other information. For this reason, there is always an asymmetry between what is predicted and what is being used to make the prediction. The predictors and the predicted quantity might be different both in number and character. This asymmetry is evident even in the case of predicting one test score, **Y**, from another, **X**. In this simplest case, it has been known since the 19th century that the usual linear regression function for predicting **Y** from **X** is not the inverse of the linear regression function for predicting **X** from **Y** (Galton, 1888). This is a basic aspect of

the asymmetry between the predictor score and the predicted score. It is highlighted in requirement (c) of Section 2.4.1.

If **X** and **Y** denote the scores on the two tests for examinees who are from a population, **P**, then denote the conditional expectation (or conditional mean/average) of **Y** given **X** over **P**, by

$$\mathrm{E}(\mathbf{Y} \mid \mathbf{X} = x, \mathbf{P}). \tag{2.1}$$

This conditional expectation is a standard method for predicting **Y** from **X**. If **X** has the value x, then the equation $y = \mathrm{E}(\mathbf{Y} \mid \mathbf{X} = x, \mathbf{P})$ predicts y to be the value of **Y**. The prediction of **Y** from **X** is an example of a direct link between the scores on the two tests.

Unless **Y** is functionally dependent on **X**, there is always some amount of *error* or uncertainty in any prediction. The error in this prediction is how far $\mathrm{E}(\mathbf{Y} \mid \mathbf{X} = x, \mathbf{P})$ is from the actual value of **Y**; that is, the difference

$$\mathbf{Y} - \mathrm{E}(\mathbf{Y} \mid \mathbf{X} = x, \mathbf{P}). \tag{2.2}$$

The conditional expectation is the *best* predictor of **Y** in the sense that any other predictor of **Y** from **X**, say $y = \mathrm{Pred}(x)$, will have a larger expected squared error in expression (2.2); that is,

$$\begin{aligned} &\mathrm{E}\left[\left(\mathbf{Y} - \mathrm{Pred}(\mathrm{x})\right)^2 \mid \mathbf{X} = x, \mathbf{P}\right] \geq \\ &\mathrm{E}\left[\left(\mathbf{Y} - \mathrm{E}\left(\mathbf{Y} \mid \mathbf{X} = x, \mathbf{P}\right)\right)^2 \mid \mathbf{X} = x, \mathbf{P}\right] = \mathrm{Var}\left(\mathbf{Y} \mid \mathbf{X} = x, \mathbf{P}\right), \end{aligned} \tag{2.3}$$

as shown in Cramér (1946), Parzen (1960), and others.

The conditional variance in Equation 2.3 is also called the conditional *prediction error* in the context of predicting **Y**-scores from **X**-scores. Other types of predictor or prediction method minimize other measures of prediction error, a subject too large for us to do much more than merely mention. For example, see Blackwell and Girshick (1954), Parzen (1960), or the discussion of best linear predictors in Holland and Hoskens (2003).

Using regression methods, both the conditional expectation, $\mathrm{E}(\mathbf{Y} \mid \mathbf{X} = x, \mathbf{P})$, and the conditional prediction error can be estimated from data in which examinees are sampled from **P** and tested with both **X** and **Y**. Discussions of regression methods are so widely available that no more details are given here about the variety of possibilities; for example, see Moore and McCabe (1999) or Birkes and Dodge (1993).

An appropriate use of predicting to make a link between two tests arises when an examinee's score on one test is used to predict how he or she will

perform on another test. An example is the use of PSAT/NMSQT® scores to forecast how an examinee will perform on the SAT® a year or so later. For example, periodically a year's worth of SAT data from students who have taken both tests is used to estimate the conditional distribution of SAT scores given the corresponding (verbal or mathematical) PSAT/NMSQT score (see Educational Testing Service, 1999). This conditional distribution predicts the range of likely performance on the SAT given an examinee's PSAT/NMSQT score. If these predictions are applied to examinees who are similar to those in the population from which the prediction equations are derived, then they are likely to be useful. For examinees who are very different from those whose data were used to estimate the conditional distributions, these predictions are less likely to be accurate.

2.2.2. Projecting Distributions of Observed Scores

Related to predicting individual scores on a test is the problem of *projecting distributions of scores* on one test from those on another test. In this case, as described earlier for predicting a score on **Y** from a score on **X**, data obtained from a sample of examinees who take both **X** and **Y** is used to estimate the conditional distribution of **Y** given **X** on a particular population, say **P**. Denote the conditional cumulative distribution function (cdf) of **Y** given $\mathbf{X} = x$ in **P** by

$$\Pr\{\mathbf{Y} \le y \mid \mathbf{X} = x, \mathbf{P}\}. \tag{2.4}$$

The data can be used to estimate the cdf in Equation 2.4. Now suppose that in another population, say **Q**, there are data for the distribution of **X**, but not for **Y**. If the distribution of **X** in **Q** is somewhat different from that of **X** in **P**, it might be desired to *project* the distribution of **X** in **Q** to obtain an estimate of the cdf of **Y** in **Q**, $F_{\mathbf{YQ}}(y)$, using methods that are based on the formula

$$F_{\mathbf{YQ}}(y) = \Pr\{\mathbf{Y} \le y \mid \mathbf{Q}\} = \mathrm{E}[\Pr\{\mathbf{Y} \le y \mid \mathbf{X}, \mathbf{P}\} \mid \mathbf{Q}]. \tag{2.5}$$

In Equation 2.5, the outer expectation (or averaging) is over the distribution of **X** in **Q**. Strictly speaking, Equation 2.5 is valid only if the conditional distribution of **Y** given **X** is the same in both **P** and **Q**; that is, if

$$\Pr\{\mathbf{Y} \le y \mid \mathbf{X} = x, \mathbf{P}\} = \Pr\{\mathbf{Y} \le y \mid \mathbf{X} = x, \mathbf{Q}\}. \tag{2.6}$$

Equation (2.6) is a type of *population invariance* assumption because it requires the conditional distribution that holds for one population to also hold for another population. Assumptions that are identical to Equation 2.6 also arise in various cases of scaling and equating. Population invariance

assumptions, like Equation 2.6, pervade all aspects of scaling and equating where there are missing data in the sense that in the above example the data for **Y** in **Q** are missing.

An important example of projecting a score distribution arises when **X** and **Y** are both given to a sample of examinees in Year 1, and then in Year 2, only one of them, say **X**, is given. To predict what the distribution of **Y** would have been had it also been given in Year 2, projection methods provide a way of doing this. They are based on Equation 2.5, with **P** representing the data from Year 1 and **Q** representing the data in Year 2. The need for the population invariance assumption in Equation 2.6 is quite evident in this example.

Pashley and Phillips (1993) provided an example of projecting scores from the International Assessment of Educational Progress (IAEP) to the scale of the National Assessment of Educational Progress (NAEP). Williams, Rosa, McLeod, Thissen, and Sanford (1998) gave a detailed discussion of an example of projecting scores from a state assessment to the NAEP scale, which is the focus of the chapters by Braun and Qian (Chapter 17), Koretz (Chapter 18), and Thissen (Chapter 16).

So far, the discussion has concerned only prediction methods that directly link observed scores on the tests to each other. There are other forms of prediction worthy of mention for completeness (e.g., methods that use observed scores to predict *true scores*).

2.2.3. Predicting True Scores

The oldest version of predicting true scores from observed scores is Kelley's formula that predicts the true score on **Y** from the observed score on **Y** (Kelley, 1927). This idea was generalized in Wainer et al. (2001) to the prediction of true scores on one test from the observed scores on it and some other tests. They referred to the predicted true scores as *augmented scores*. Holland and Hoskens (2003) considered the problem of predicting true-scores from observed scores where the true-scores come from one test, **Y**, and the observed scores come from another test, **X**. They showed that the usual linear regression function, which predicts the observed scores of **Y** from the observed scores of **X**, is an appropriate predictor of the true score of **Y**, but that the usual measure of prediction error from linear regression is too large and needs to be adjusted by the reliabilities of the two tests.

2.2.4. Summary

It was recognized very early that prediction methods were not satisfactory ways of creating *comparable scores*, as the early forms of scale aligning were called. Thorndike (1922) and Otis (1922) gave the first arguments for why linear regression was not a satisfactory method of finding comparable scores. Later, Flanagan (1951) emphasized the lack of symmetry of regression functions, thereby connecting regression methods to the failure to satisfy requirement (c) of Section 2.4.1. The distinction between prediction and equating has been repeatedly reaffirmed over the years; see Hull (1922), Flanagan (1939, 1951), Lord (1950, 1955, 1982), Angoff (1971), Mislevy (1992), Linn (1993), and Holland and Dorans (2006).

2.3. Scale Aligning

The methods of aligning scales are the second oldest group of linking methods. The need to make scores on different tests comparable (i.e., scaling) and the invention of methods to do it has a history almost as old as the field of psychometrics itself. Procedures for scaling were initially called methods for creating comparable scores. Kelley (1914) discussed problems with the methods proposed in Starch (1913) and modified in Weiss (1914) and Pinter (1914) for putting into comparable units the Ayers and the Thorndike methods of scoring of handwriting. Pinter had a sample of handwriting from examinees who had been judged using both methods. Weiss advocated setting the means of the scores on both tests equal to 50 by a multiplicative factor. Kelley showed that this method could give absurd results in various circumstances and proposed, instead, to use standard scores as comparable measures (i.e., to subtract the mean and divide by the standard deviation of each measure). Using standard scores to scale tests has been used widely since that time. Treating standard scores as equivalent leads to the method of linear equating. Kelley explicitly titled his article "Comparable Measures" and used the terms *equate* and *equating* to refer to the results of setting comparable scores equal.

The influential textbook by Kelley (1923) had a chapter titled "Comparable Measures" in which he (a) again showed that the method proposed by Weiss (1914) can lead to absurd results, (b) asserted that Galton had, decades earlier, used a version of standard scores to compare quantities that are measured on different scales, (c) advocated standard scores and showed that they equal the ratio method only when special conditions hold, and (d) discussed the equal successive percentiles method to define comparable scores; this is an early form of equipercentile

equating (Equation 11 in Kelley). Kelley referred to even earlier uses of the equal successive percentile method in Otis (1916, 1918).

These references suggest that by the time of the US entry into World War I, those who worked with test data had some familiarity with both the linear and the equipercentile methods of scaling the scores from different tests. von Davier, Holland, and Thayer (2004b) quoted Kelley (1923) to indicate that he was aware of the dual influence of examinee ability and test difficulty on test scores and this needed to be accounted for in scaling tests.

The goal of scale aligning is to transform the scores from two different tests onto a common scale. Scaling transformations take scores from two different tests, **X** and **Y**, and put them onto a common scale. Such aligned scales imply an indirect linking of the scores on **X** and **Y**. More specifically, the implied linking is found by taking a score on **X**, transforming it to the common scale, and then inverting the **Y**-to-scale transformation to find the corresponding value for **Y**. The result is an indirect link from scores on **X** to those on **Y**. All methods of scale aligning can create indirect links between tests in this way.

It should be emphasized that although the implied indirect links always exists, their meaningfulness depends on many factors, and the indirect link is rarely the main purpose for putting **X** and **Y** onto a common scale.

The subcategories of scaling form a continuum starting with situations where the tests measure different constructs all the way to those where the tests measure similar constructs. The next five subsections briefly describe the six types of scaling along this continuum. Figure 2.3 illustrates the subcategories within the overall linking category of scale aligning.

2.3.1. Battery Scaling: Different Constructs and a Common Population of Examinees

When two or more tests that measure different constructs are administered to a common population, the scale scores for each test can be transformed to have a common distribution for this population of examinees (i.e., the *reference population*). Kolen (2004a) denoted this case as *battery scaling*. Battery scaling has been used for many years. Flanagan (1951) described it in an educational testing context, but its roots can be traced back at least to Kelley (1914), where the scores on the different tests were given the same mean and variance in the reference population. Kelley (1923) and Angoff (1971) referred to scores from tests that measure different constructs but that are scaled so that they have the same distributions on a common population as *comparable measures* (Kelley, 1923) or *comparable scores* (Angoff, 1971).

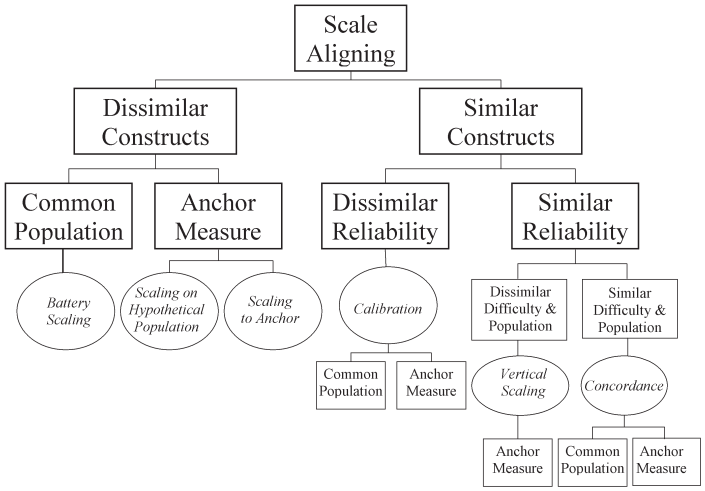

Figure 2.3. The types of linking method within the overall linking category of *scale*.

The data collected for battery scaling is usually either (a) a sample of examinees, all of whom take all of the tests, or (b) several equivalent (i.e., random) samples of examinees from a common population who take one or just some of the tests. In this way, all of the tests are taken by equivalent groups of examinees from the reference population. Thus, for each test being scaled, **Y**, the data can be used to estimate the cdf of **Y** over the reference population, **P**; that is,

$$F_{\mathbf{YP}}(y) = \Pr\{\mathbf{Y} \le y \mid \mathbf{P}\}. \tag{2.7}$$

Y is then put on the common scale by a transformation of the form

$$\mathrm{s} = S(F_{\mathbf{YP}}(y)), \tag{2.8}$$

where $S(u)$ is an arbitrary *scaling function* selected to give the scaled version of **Y** a particular distributional form. A common example of such a scaling function is the inverse of the Normal or Gaussian distribution so that the distribution of the scaled scores is approximately Gaussian (Kolen & Brennan, 2004).

The value of making the scales of different tests comparable in this special sense is that examinees will correctly interpret differences in the scores across the battery of tests. A higher score on one test will indicate better performance on that test when compared to a lower score on another test (relative to the population **P**). Effectively, comparing scaled scores

becomes the same as comparing percentiles in the reference population when the scales have been aligned this way. Measures or scores on comparable scales could be useful for comparing the strengths and weaknesses of examinees who are similar to those in the reference population. For examinees who are different from those in the reference population, such interpretations might not be as useful.

Although the scales on the different tests are made comparable in this special sense, the tests measure different constructs. The implied indirect link between the scores on the different tests, described earlier, can be used to indicate comparable performance on the different tests (relative to the reference population), but it has no meaning as a way of transforming a score on a test of one construct into a score that is an appropriate measure for another construct.

The recentering of the SAT scale is an example of battery scaling (Dorans, 2002). The scales for the SAT-verbal (SAT-V) and SAT-mathematical (SAT-M) scores were redefined so as to give the scaled scores on the SAT-V and SAT-M the same distribution in a reference population of students tested in 1990. The redefined score scales replaced the original score scales, which had been defined for a reference population tested in 1941. The new score scales enable a student whose SAT-M score is higher than his SAT-V score to conclude that he/she did in fact perform better on the mathematical portion than on the verbal portion, at least in relation to the students tested in 1990. When the scales of tests are not aligned in this way, such inferences are not necessarily accurate. As the population of students taking the SAT becomes less like the reference population tested in 1990, the simple interpretation of better performance on one test compared to another, based solely on the scaled scores, will become less accurate. Finally, it should be obvious that the indirect link between the SAT-M and SAT-V has no meaning as a way of turning a score on one of these tests into a score on the other.

2.3.2. Anchor Scaling: Different Constructs and Different Populations of Examinees

An important approximation to battery scaling arises when two or more tests that measure different constructs are administered to *different* populations and a common measure (the *anchor measure*) is available for all of the examinees in these different populations. *Anchor scaling* refers to this general class of scaling method. Mislevy (1992) and Linn (1993) used the term *statistical moderation* to refer to cases of anchor scaling.

In the typical application of anchor scaling, it is possible for one or more of the tests being scaled to be completely inappropriate for the examinees

taking some of the other tests. Language examinations provide good examples of this: A test of French is inappropriate for examinees who are unfamiliar with French. In other situations, examinees might choose which test to take based on the courses they have taken in school. Because of these selective factors, the samples of examinees taking the different tests are usually *not equivalent*, and the anchor measure is the information used to both measure and to adjust for this. Anchor scaling necessarily involves incomplete test data because some tests are given to certain subgroups of examinees, but not to all of them. Anchor scaling is an *approximation* to battery scaling because of the potential inequivalence of the samples of examinees taking each of the tests. In contrast, when different samples of examinees take different tests for battery scaling, these samples are designed to be *equivalent* samples of examinees.

The inequivalence of the samples used in anchor scaling requires the scaling methods used to make assumptions about the anchor measure that are not easily evaluated. The more strongly the anchor measure is related to the different tests being put on a common scale, the more satisfactory the resulting scale alignment will be, but other than that, little more can be said in general.

2.3.2.1. Scaling on a Hypothetical Population

There are two distinct ways that the anchor measure is used in anchor scaling. The first approach is very similar to projecting score distributions, discussed in Section 2.2. This approach has no commonly accepted name, so Holland and Dorans (2006) proposed identifying it as *scaling on a hypothetical population* (SHP). To outline this approach and to relate it to projecting score distributions, suppose that **Y** denotes a test to be scaled and **A** is the anchor measure. The data for the examinees taking **Y** and **A** are used to estimate the conditional distribution of **Y** given **A** in the population of examinees (denoted by $\mathbf{P}_Y$) who take test **Y**. As indicated earlier, $\mathbf{P}_X$ and $\mathbf{P}_Y$ might be different for different tests, **X** and **Y**. As in Section 2.2, denote the cdf of this conditional distribution by

$$\Pr\{\mathbf{Y} \le y \mid \mathbf{A} = a, \mathbf{P}_Y\}. \tag{2.9}$$

Next, this estimated conditional distribution is averaged over a hypothetical distribution for **A**, the distribution of **A** in the *hypothetical* population, **P**, to obtain an estimate of the cdf of **Y** in the hypothetical **P**; that is,

$$\Pr\{\mathbf{Y} \le y \mid \mathbf{P}\} = \mathrm{E}[\Pr\{\mathbf{Y} \le y \mid \mathbf{A}, \mathbf{P}_Y\} \mid \mathbf{P}]. \tag{2.10}$$

In Equation 2.10, the outer expectation is over the distribution of **A** in the hypothetical population. These cdfs are found for each of the tests

being scaled. The estimated cdf for **Y** on the hypothetical population, defined in Equation 2.10, is then treated as if it is *the* cdf of **Y** on a common population. Once this is done, the problem is regarded as the simpler case of battery scaling and the same scaling techniques are used from that point forward.

As in the case of projection in Section 2.2, in order for Equation 2.10 to hold, a population invariance assumption, similar to Equation 2.6, must hold. The weaker the correlation between the anchor measure and the test, the less likely it is for this population invariance assumption to hold, even approximately.

It should also be pointed out here that there is nothing in the above analysis that requires the anchor measure to be a single score or number; it could involve more than one score, as the next example illustrates.

The construction of the hypothetical population is critical to the success of this method because the linking is population dependent. Although a variety of hypothetical populations might be posited in a particular setting, they are unlikely to be equally plausible. Great care needs to be exercised in the construction of the population.

An example of SHP is given by the scaling of the various subject area tests of the SAT. Typically, students take the SAT, and then some of them might take one or more subject tests. All of these scores are then presented as part of their college admissions materials, and the results of the subject tests for different examinees are treated as if they are on comparable scales. In this application, the SAT-V and SAT-M scores are used as the anchor measures. The hypothetical population is taken to be the population on which the SAT-V and SAT-M scales were established. SHP is closely related to poststratification equating, mentioned in Section 2.4.

2.3.2.2. Scaling to the Anchor

The second approach to anchor scaling also has no commonly accepted name, so Holland and Dorans (2006) identified it as *scaling to the anchor* (STA). In this approach, the data for the examinees taking test **Y** are used to estimate a function linking scores on **Y** to those on **A** using the data from $\mathbf{P}_Y$. This is done for each of the tests to be scaled and these linking functions are used to put each of the tests onto the scale of the anchor measure. Strictly speaking, in order for STA to be valid, the estimated linking functions for each test should not depend on the choice of the population used for each linking. This is a population invariance assumption similar those mentioned in Section 2.4.3 for chain equating.

Linn (1993) indicated that the STA approach was used to bring comparability to scores on tests that are specific to particular schools in a school district. The anchor measure is a common districtwide examination

score, and the scores from the locally developed tests in each school are put on a common scale using the STA approach to anchor scaling.

One difference between STA and SHP is that for STA, the measure needs to be a single score or number, whereas, indicated earlier, the SHP can operate on multiple sets of scores. See McGaw (1977) and Keeves (1988) for more discussion of STA, where it was referred to as an example of *moderation*.

2.3.3. Vertical Scaling: Similar Constructs and Similar Reliability, But Different Difficulty and Different Populations of Examinees

Tests of academic subjects targeted for different school grades might be viewed as tests of similar constructs that are intended to differ in difficulty—those for the lower grades being easier than those for the higher grades. It is often desired to put scores from such tests onto a common overall scale so that progress in a given subject can be tracked over time. This type of scaling is called *vertical scaling* (Kolen & Brennan, 2004). It has been called other things as well. For example, Angoff (1971) called it *calibrating tests at different levels of ability* and the term *vertical equating* is also used.

A topic, such as mathematics or reading, when considered over a range of school grades, has several subtopics or dimensions. At different grades, different aspects, or dimensions, of these subjects are relevant and tested. For this reason, the constructs being measured by the tests for different levels might differ somewhat, but the tests are often similar in reliability.

Vertical scaling shares some features with anchor scaling (Section 2.3.2). In particular, the tests to be scaled are, to some degree, inappropriate for all but one or a few grades, so the samples of examinees who take each test are not equivalent in the sense that they are for battery scaling (Section 2.3.1). However, due to the range of ages and grades that are usually involved, there is rarely an appropriate anchor measure that is available for every examinee. Instead, the tests given to neighboring grades might share some common material that can serve as an anchor test that connects a pair of tests for different grade levels but not all of the tests being scaled. This common material will be different for different pairs of tests given to neighboring grades. Methods such as SHP and STA, described briefly in Section 2.3.2, might be used to put the tests given to neighboring grades onto a common scale, and these can then be connected up to form an overall scale for the entire vertical system of tests. Item response theory (IRT) is also used to link these scales. See Kolen and Brennan (2004), Petersen, Kolen, and Hoover (1989), and the chapter by

Kolen (2006) for more discussions of these and other methods used in vertical scaling.

There is usually a close connection between the material tested in a given test and the curriculum for that grade. For this reason, vertical scaling might be sensitive to population differences, such as school grade or age. For example, scaling a fourth-grade reading test to a fifth-grade reading test on a sample of fifth graders is likely to disagree somewhat with the link obtained from a sample of fourth graders. For more discussion of these issues, see the chapter by Kolen (2006), as well as Harris, Hendrickson, Tong, Shin, and Shyu (2004), Hoover, Dunbar, and Frisbie (2001), and Kolen (2003). Chapters by Harris (Chapter 13), Patz and Yao (Chapter 14), and Yen (Chapter 15) discussed issues in vertical scaling in depth. For an illustration of vertical scaling, see Williams. Pommerich, and Thissen (1998).

Vertical scaling can be viewed as producing indirect links between the scores on the different levels of the tests, but these links are of less interest than the comparisons of scores on the same scale for the same student on the different tests in order to measure his or her learning and growth.

2.3.4. Calibration: Same Construct, Different Reliability, and the Same Population of Examinees

Kolen and Brennan (2004) indicated that in the test-linking literature, the term *calibration* is used in a variety of senses. In Angoff (1971), it referred to vertical scaling (Section 2.3.3). In Petersen et al. (1989), *calibration* referred to the estimation of item response theory (IRT) item parameters so that they were on a common scale. This usage is standard in the IRT literature (Lord, 1980; Thissen & Wainer, 2001; Yen & Fitzpatrick, 2006). In Linn (1993), *calibration* referred to methods of score linking for tests that measure the same constructs but that have different statistical characteristics—in particular, different reliability *or* difficulty.

Here the term *calibration* is used to refer to situations in which the tests measure the same construct, have similar levels of difficulty, but differ in reliability (usually test length). To add to the confusion, Angoff (1971) regarded this use of calibration as an example of *equating* tests of differing reliability; in this framework, *equating* is reserved for tests of equal or at least very similar reliability. The classic case of calibration in the sense used here is scaling the scores of a short form of a test onto the scale of its full or long form.

For calibration, there might be some ambiguity as to whether the linking is direct or indirect. The short form is often derived from the long form so that it usually makes more sense to scale from the less reliable test to the

more reliable one than vice versa. It is intuitively obvious as well that simply putting the scores of the short form onto the scale of a more reliable long form cannot increase the actual reliability of the short form.

2.3.5 Concordances: Similar Constructs, Difficulty, and Reliability

Sometimes the tests to be linked all measure similar constructs, but they are constructed according to different specifications. In most cases, they are similar in test length and reliability. In addition, they often have similar uses and might be taken by the same examinees for the same purpose. The use of the linking is to add value to the scores on both tests by expressing them as if they were scores on the other test. Concordances represent scalings of tests that are very similar but that were not created with the idea that their scores would be used interchangeably. See Pommerich and Dorans (2004a) for a thorough discussion of many aspects of concordances.

Many colleges and universities accept scores on either the ACT® or SAT for the purpose of admissions decisions, and they typically have more experience interpreting the results from one of these tests than the other. Dorans, Lyu, Pommerich, and Houston (1997) reported a concordance table or function that linked the scores on each of these two tests to each other. This concordance was based on data from more than 100,000 examinees who had taken both tests within a restricted time frame. If their applicants were not widely different from those in this large sample, this concordance enabled admissions officers to align cut-scores on these two similar but somewhat different tests better than they could have using the limited data typically available to them.

Because the tests being linked measure somewhat different constructs and are constructed in different ways, concordances are potentially sensitive to the population of examinees whose data are used to estimate the concordance function. Dorans and Holland (2000) and Holland and Dorans (2006) argued that when the data indicate that substantially different concordance functions hold for large subpopulations of examinees (e.g., males and females), separate concordance functions ought to be considered for these groups, lest one group be disadvantaged by the use of a pooled concordance function for all. Dorans (2004d) discussed this point for the ACT and SAT. In practice, separate concordances might not be feasible for a variety of reasons, including a perceived unfairness in high-stakes uses of the tests.

Concordances are examples of scalings that produce direct links between the scores on the two tests.

Chapters by Pommerich (Chapter 11), Sawyer (Chapter 12), and Dorans and Walker (Chapter 10) addressed concordances in more detail. The chapters by Brennan (Chapter 9), Eignor (Chapter 8), and Liu and Walker (Chapter 7) addressed linking issues for testing programs in a state of transition, either with regard to mode of administration or test content. These linkages might be concordances, calibrations, or equatings.

2.4. Equating: Same Construct and the Same Intended Difficulty and Reliability

Equating is the third category of linking methods in this framework. All linking frameworks define equating as the strongest form of linking between the scores on two tests. In this chapter, equating represents the end point of a continuum that begins with methods that make no assumptions about the relationships between the tests being linked (prediction and battery scaling) and proceeds to methods that are appropriate for linking tests that are very similar (concordances and equating). Equating might be viewed as a form of scaling in which very strong requirements are placed on the tests being linked.

The purpose of equating is to allow the scores from each test to be used interchangeably, as if they had come from the same test. This purpose puts strong requirements on the two tests and on the method of linking. Among other things, the two tests must measure the same construct at similar levels of difficulty and reliability.

The earliest example of equating alternative forms of the same tests is not known to this author, but there is an early example of alternative forms that *were not equated*: the Army Alpha Test used by the American army during World War I. By the end of 1918, the army had tested over 1.7 million men using the Alpha and Beta. The Alpha was targeted for examinees who could read and write English and the Beta was for those who could not. Yoakum and Yerkes (1920) gave a detailed description of both instruments. They indicated that the Alpha had five different test forms: "To avoid . . . the risk of coaching, several duplicate forms of this examination have been made available" (p. 18). Thus, by this early date, test security issues had already led to the use of alternate forms, at least for the Alpha. Yoakum and Yerkes said little about how the alternate forms of the Alpha were constructed, but the following passage suggests that they used random assignment of test items to forms to help ensure the similarity of the alternate forms. "All five forms of the group examination were used in the pre official trial of the tests. The differences in forms were so slight as to indicate the success of the random method of selecting items" (p. 8).

Under appropriate conditions, assigning test items to forms at random will produce nearly parallel test forms that are similar but not identical in difficulty. In the next sentence, Yoakum and Yerkes indicated that the five forms were not exactly equivalent: "Form B proved more difficult than the other forms" (p. 8).

Nothing more is said about the issue of Form B's difficulty, and in all probability, scores on the different forms of the Alpha were treated as sufficiently similar so that they were not equated, even though the linear and equipercentile methods for doing so were known and available by that time.

Of greater concern to the army statisticians was the comparability of scores achieved on the Alpha and Beta versions of the test. A special sample of military personnel was tested with both, and these data were used to put the Alpha and Beta on a common 7-point scale (A, B, C+, C, C−, D, D−). Because these two tests were quite different in terms of format and questions asked, this was a case of battery scaling rather than of test equating. Indeed, Thorndike (1922) referred to three distinct scalings of the Alpha and Beta.

The example that Kelley criticized in 1914 was also a form of battery scaling rather than equating. The two methods of assessing handwriting were very different scoring methods and would not, in current terminology, be construed to be alternative forms of the same test. The problem that interested Pinter (1914) and Starch (1913) was to measure the accuracy/stability of these different handwriting measures. Kelley referred to an earlier work by Woodworth (1912), which used standard scores to combine the results of several tests. Otis (1918) was also interested in the problem of combining test results when the tests were on quite different subjects: spelling, arithmetic, synonyms, proverbs, and so forth. Thus, these early uses of comparable scores were not to equate scores in the sense used here, but, rather, as intermediate battery scalings needed to solve other problems.

Terman and Merrill (1937) discussed their revised edition of the Stanford-Binet test. Two alternative forms of the new edition were produced, but they were not equated directly. Rather, both were treated separately and the scores of each one put on the IQ scale using battery-scaling methods. In the next edition of the Stanford-Binet test, the second form was eliminated because it was rarely used.

Thus, the need, or at least the desire, to equate scores on alternate forms of the same test probably arose decades after the invention of scaling methods and of the two standard methods for equating: the linear and equipercentile methods. In 1938 two forms of the College Board's SAT tests were given in the same year, and the need to equate them became evident by 1940. Early versions of anchor-test equating were used to

remove the effect of differential form difficulty for the SATs in 1941. In 1942 the SAT verbal and math scales were linked back to the verbal scale established in April 1941; all linkings subsequent to 1942 were equatings (Donlon & Angoff, 1971; Dorans, 2002). Lord (1950, 1955) credited Ledyard R Tucker with devising the anchor-test methods used to equate the SATs during the 1940s; these methods, in various versions, continue to be used.

Test equating is a necessary part of any testing program that continually produces new test forms and for which the uses of these tests require the meaning of the score scale be maintained over time. Although they measure the same constructs and are usually built to the same test specifications or test blueprint, different editions or forms of a test almost always differ somewhat in their statistical properties. For example, one form might be harder than another, so without adjustments, examinees would be expected to receive lower scores on the harder form. A primary goal of test equating for testing programs is to eliminate the effects on scores of these unintended differences in test form difficulty. For many testing programs, test equating is necessary to be fair to examinees taking different test forms and to provide score-users with scores that mean the same thing, regardless of the tests taken by examinees (Angoff, 1971; Kolen & Brennan, 2004; Petersen et al., 1989).

In testing programs with high-stakes outcomes, it cannot be overemphasized how important it is that test equating be done carefully and accurately. The released scores are usually the most visible part of a testing program, even though they represent the end point of a long test production, administration, and scoring enterprise. An error in the equating function or score conversion function might change the scores for many examinees. The credibility of testing organizations has been called into question over test equating problems, in ways that rarely occur when, for example, flawed test questions are discovered in operational tests. Chapters 5, 6, and 4 by Cook, von Davier, and Petersen, respectively, in this volume addressed issues related to equating.

2.4.1. What Makes a Linking an Equating?

All forms of test score linking involve some of the same ingredients. These include (a) two or more tests and rules for scoring them, (b) scores on these tests from one or more samples of examinees, (c) an implicit or explicit population of examinees on which linking takes place, and (d) one or more methods of estimating or calculating the linking function. What distinguishes test equating from other forms of linking is its demanding goal of allowing the scores from both tests to be used interchangeably for any purpose.

In the context of a testing program that continually produces new test forms that are required to produce scores on the same scale, test equating is often regarded as the first part of a two-step process by which scores on new tests are put onto the reporting scale. The first step is the computation of the *equating function*, $y = e(x)$, that links the raw scores on a new test, **X**, to those of an old test, **Y**—the so-called *raw-to-raw equating*. The second step is the conversion of these equated **X** raw scores to the reporting scale. In practice, there is an old form conversion function that maps the raw scores of the old test, **Y**, to the scale, call it $S = s(y)$. The old form conversion function is composed with the equating function, $e(x)$, to put the raw scores of **X** onto the reporting scale; that is, the new form conversion function is $s(e(x))$.

An alternative approach is to use the methods of IRT to find a direct conversion of **X**-scores to the common IRT scale rather than going through an old test, **Y**. This method, in principle, does not even require an old test, but could involve portions of several old tests. Discussion of this approach is beyond the scope of this chapter. Instead, the focus here is on equating functions.

Dorans and Holland (2000) outlined five requirements that are widely viewed as necessary for test equating to be successful. The order in which these requirements are listed corresponds roughly to the order of their appearance in the literature.

a. *The equal construct requirement*: The tests should measure the same constructs.
b. *The equal reliability requirement*: The tests should have the same reliability.
c. *The symmetry requirement*: The equating function for equating the scores of **Y** to those of **X** should be the *inverse* of the equating function for equating the scores of **X** to those of **Y.**
d. *The equity requirement*: It should be a matter of indifference to an examinee to be tested by either one of two tests that have been equated.
e. *The population invariance requirement*: The choice of (sub)population used to estimate the equating function between the scores of tests **X** and **Y** should not matter; that is, the equating function used to link the scores of **X** and **Y** should be *population invariant*.

Both formal and informal statements of subsets of these five requirements appeared in a variety of earlier sources, including Lord (1950), Angoff (1971), Lord (1980), Petersen et al. (1989), and Kolen and Brennan (2004). Kolen (Chapter 3, Section 3.2) pointed out the importance

of common conditions of measurement as well as common content as a requirement for equating.

In practice, requirements (a) and (b) mean that the tests need to be built to the same specifications and administered under the same conditions of measurement, whereas requirement (c) precludes regression methods for predicting **Y**-scores from **X**-scores from being a form of test equating. Lord (1980) indicated that requirement (d) explains why both requirements (a) and (b) are needed. Requirement (d) is, however, hard to evaluate empirically and its use is primarily theoretical (Lord, 1980; Hanson, 1991). Furthermore, requirement (e), which is easy to use in practice, also can be used to explain why requirements (a) and (b) are needed (Holland & Dorans, 2006). Dorans and Holland (2000) used requirement (e) to develop quantitative measures of equitability. Their measures indicate the degree to which equating functions depend on the subpopulations used to estimate them.

The other cases of score linking are likely to violate at least one of the five requirements for equating. Concordances are used with tests that measure similar but different things and do not share common test specifications. Although they might have a similar difficulty and reliability, they will satisfy requirement (a) only approximately and this might be detected by the failure of requirement (e) and possibly requirement (d). Tests that are vertically scaled might be on such different aspects of a school subject that requirement (a) is not satisfied, at least when the gap between the grades is large and the differences in difficulty might be so great that, regardless of attempts to scale them appropriately, examinees will definitely prefer one test over the other, thus violating requirement (d) and probably requirement (e) as well. Calibrating a short form to a long form violates requirement (b) and is likely to violate requirements (d) and (e).

The tests that are scaled by either battery scaling or anchor scaling are usually measures of different constructs by design so that requirement (a) is not satisfied. Furthermore, scaling tests of different constructs will also tend to fail to satisfy requirements (d) and (e) for important subgroups of examinees. The direct and indirect linkings that arise in scaling are invertible, so requirement (c) is usually satisfied.

Finally, prediction methods need not satisfy any of the five requirements. The asymmetry between predictors and outcomes violates requirement (c). Furthermore, requirements (a) and (b), measuring the same construct and being equally reliable, affect only the quality of the prediction; less related or less reliable tests make poorer predictors of the scores on another test. Requirement (d) plays no role in prediction. Finally, it often makes sense to include subgroup membership as predictors to improve prediction. This incorporates population sensitivity directly into

the prediction, whereas equating functions should not depend on subpopulations, according to requirement (e).

The difference between prediction and equating has been pointed out repeatedly over the last century. To give an example that shows how test equating and predicting can work together but do different things, suppose the scores from one testing program are used to predict some outcome variable, such as first-year college grades, using regression methods. In this case, the test score is being used as a predictor. It is routine to use the equated scores that come from different test forms as interchangeable values of the predictor. The predictions benefit from a prior test equating because test equating eliminates the need to distinguish between the scores on the various forms of the test that are used as predictors. This application occurs every time test scores from a testing program are used as predictors in validity studies. However, the predicted average grades from the test score would never be construed as an *equating* of test scores and first-year grades.

2.4.2. A Crucial Consideration for Scale Aligning and Equating

There is one common concern for all of the methods that are grouped under categories of scale aligning and equating. Appropriate attention must be given to the control of differential examinee ability in the linking process. To be clearer about this, suppose that two different tests are given to two different groups of examinees. In the two distributions of resulting scores, there are two ever-present factors that can influence the results, regardless of how similar the score scales of the tests appear. One is the relative *difficulty* of the two tests (which is what test scaling and equating is concerned about) and the other is the relative *ability* of the two groups of examinees on these tests (which is a confounding factor that should be eliminated in the linking process). In scaling and equating, the interest is in adjusting for differences in test characteristics and in controlling for possible examinee differences in ability when making these adjustments.

There are two distinct ways that the separation of test difficulty and differential examinee proficiency is addressed in practice. The first is to use a common population of examinees and the other is use an anchor measure. These approaches were mentioned in the discussion of scaling aligning in Section 2.3. Using the same examinees explicitly controls for differential examinee ability (i.e., they are the same examinees and have the same proficiencies). A variant of the use of a common set of examinees is to use two equivalent samples of examinees from a common population. On the other hand, when it is not possible to have samples of examinees from the same population, their performance on an anchor measure or set

of common items can quantify the differences between two distinct, but not necessarily equivalent, samples of examinees. The use of an anchor measure leads to approaches that can be more flexible than the use of common examinees (Holland & Dorans, 2006).

2.4.3. A Brief Outline of Equating Methods

Numerous methods have been developed over the years for scaling and equating tests. In the next two subsections they are organized according to whether the data collection design involves a common population or common items. The focus here is on *observed-score procedures* that directly transform (or link) the scores on **X** to those on **Y**, because these methods are the most directly related to the estimation of equating functions. True-score methods are mentioned in passing. Kolen (Chapter 3, Section 3.5) provided a more extensive consideration of methods and data collection designs.

Figure 2.4 organizes the subcategories within the overall linking category of equating.

2.4.3.1. Procedures for Equating Scores on a Common Population

Holland and Dorans (2006) discussed three data collection designs that make use of a common population of examinees: the single group (SG), the equivalent group (EG), and the counterbalanced (CB) designs. They all involve a single population, **P**. Most of this section applies easily to both the EG and SG designs. The CB design is more complicated and is omitted; for more on the CB design, see Kolen (Chapter 3), von Davier et al. (2004b), Angoff (1971), and Kolen and Brennan (2004).

Several procedures have been developed for estimating equating functions using a common population. Underlying any linking method is a *target population* of examinees, following the usage in von Davier et al. (2004b). The target population is the population for which the equating function is supposed to apply. For data collection designs that use a common population, this is also the target population. In this chapter, **T** denotes the target population of examinees.

The cdf of the scores of examinees in the target population, **T**, on test **X** is denoted by $F_{\mathbf{T}}(x)$; and it is defined as the proportion of examinees in **T** who score at or below x on test **X**. More formally, $F_{\mathbf{T}}(x) = \Pr\{\mathbf{X} \leq x \mid \mathbf{T}\}$.

The equipercentile definition of *comparable scores* is that x (an **X**-score) and y (a **Y**-score) are *comparable* in **T** if $F_{\mathbf{T}}(x) = G_{\mathbf{T}}(y)$. This means that x and y have the same percentile in the target population, **T**. When the two cdfs are continuous and strictly increasing, the equation of $F_{\mathbf{T}}(x)$ and

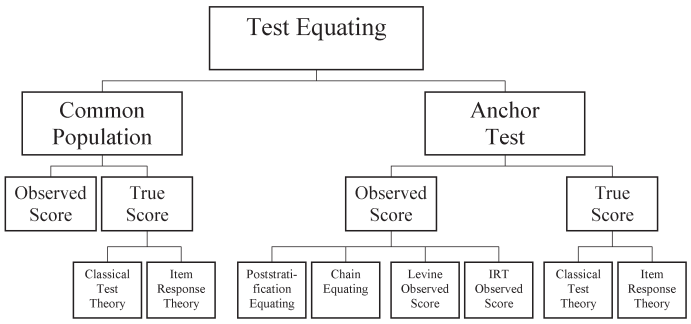

Figure 2.4. The types of linking methods within the overall linking category of *test equating*.

$G_T(y)$ can always be satisfied and can be solved for y in terms of x. This equipercentile function is used for equating, concordances, vertical scaling, battery scaling, and calibration. For equating, the influence of **T** should be small or negligible, and, in that case alone, the transformed **X**-scores are *interchangeable* with the **Y**-scores.

It is sometimes appropriate to assume that the two cdfs, $F_T(x)$ and $G_T(y)$, have the same shape and only differ in their means and standard deviations. In this case, it can be shown that the equipercentile function is the *linear linking function*. The linear linking function can also be derived as the transformation of **X**-scores that gives them the same mean and standard deviation on **T** as the **Y**-scores have.

The linear linking and equipercentile functions were introduced in the first two decades of the 20th century as methods of scale aligning. Both of these functions satisfy the symmetry requirement (c) of Section 2.4.1; that is, linking **Y** to **X** is the inverse function for linking **X** to **Y**.

The linear linking function can be viewed as the linear part of the equipercentile function (see von Davier et al., 2004b, for more details). The remainder is the nonlinear part of the equipercentile function. In the *kernel equating* method of equating (von Davier et al., 2004b), the equipercentile function and the linear linking function are shown to be two members of a two-parameter family of equipercentile functions that interpolate smoothly between these two special cases.

Although there is really only one linear linking function for the SG or EG designs, the equipercentile function can depend on how $F_T(x)$ and $G_T(y)$ are made continuous or *continuized*. Test scores are typically integers, such as number-right scores or rounded formula scores. Because

of this, the inverse function is not well defined; that is, for many values of p, there is no score y for which $p = G_T(y)$. This is not due to the *finiteness of real samples*, but, rather, to the *discreteness of real test scores*. To get around this, two methods of continuization of $G_T(y)$ are in current use.

The first is very old (Otis, 1916) and uses linear interpolation to make $G_T(y)$ piecewise linear and continuous; see Kolen and Brennan (2004). The second approach uses Gaussian kernel smoothing to continuize the discrete distributions; see Holland and Thayer (1989) and von Davier et al. (2004b). This results in a continuously differentiable $G_T(y)$. Prior to continuizing the cdfs, several authors recommended presmoothing the discrete distributions of scores (Kolen & Brennan, 2004; Kolen & Jarjoura, 1987; Livingston, 1993; von Davier et al., 2004b). In presmoothing data, it is important to achieve a balance between a good representation of the original data and smoothness. Smoothness reduces sampling variability and a good representation of the data reduces the possibility of bias.

Levine (1955) used classical test theory to derive a procedure designed to equate the true scores of **X** to those of **Y**. For a more detailed discussion of true-score equating, see Kolen and Brennan (2004). Hanson's theorem (Holland & Dorans, 2006) uses classical test theory to formalize the first four equating requirements of Section 2.4.1 and from them to derive the linear equating function as the only linear solution. Holland and Dorans also showed how Hanson's theorem shows the relationship among the linear linking function, linear regression, and true-score equating in the case of calibration (Section 2.3.4).

Lord (1980) introduced nonlinear versions of true-score equating using IRT (Kolen & Brennan, 2004).

2.4.3.2. Procedures for Linking Scores Using Common Items

The use of common items to control for differential examinee ability arises when there are two populations of examinees, **P** and **Q**, rather than just one. In this situation, **X** and a set of common items (or anchor test) **A** are taken by examinees from **P** while **Y** and **A** are taken by examinees in **Q**. Examinees take **A** and either **X** or **Y**. This is called the nonequivalent groups with anchor test or NEAT design in Holland and Dorans (2006). Kolen (Chapter 3, Section 3.5) called it the common-item nonequivalent groups design. The NEAT design is widely used because it can give greater operational flexibility than the approaches using common examinees. Examinees need only take one test, and the samples need not be from a common population.

This flexibility comes with a price, however. For one, the target population is less clearcut for the NEAT design. Which is it, **P** or **Q** or

something else? For another, the use of the NEAT design always involves making additional assumptions to allow for the missing data in the NEAT design: **X** is never observed in **Q** and **Y** is never observed in **P**.

Braun and Holland (1982) proposed that for the NEAT design, the target population be what they called the *synthetic population* created by weighting **P** and **Q**. They denoted the synthetic population by $\mathbf{T} = w\mathbf{P} + (1 - w)\mathbf{Q}$, which means that distributions (or moments) of **X** or **Y** over **T** are obtained by first computing them over **P** and **Q**, separately, and then averaging them with w and $(1 - w)$ to get the distribution or moments over **T**. The definition of the synthetic population forces the user to confront the need to create distributions (or moments) for **X** on **Q** and **Y** in **P**, where there are no data. This is why assumptions must be made about the missing data in the NEAT design.

There are three distinct sets of assumptions about the missing data that have been used to estimate observed-score equating functions for the NEAT design. These are the (a) *post-stratification equating* type, (b) *chain equating* type, and (c) *Levine* type discussed in detail in Holland and Dorans (2006). These three sets of assumptions all have the form that some aspect of the equating is the same for populations **P** and **Q**. The first two types of assumption can produce both the linear linking and equipercentile functions, whereas the Levine type, being based on classical test theory, only produces a linear function that need not be a linear linking function that describes the linear portion of the equipercentile function.

In general, the three sets of assumptions result in different equating functions; however, when $\mathbf{P} = \mathbf{Q}$, all three sets of assumptions result in the same linear or nonlinear equating functions.

For the NEAT design, there are also linear and nonlinear true-score equating functions available that use either classical test theory or IRT (Kolen & Brennan, 2004).

In the next chapter, Kolen describes various data collection designs and the methods used for equating and other types of linking in greater detail.

2.5. A Brief Note on the Theory of Equating

The theory underlying test equating has evolved slowly over the years. The methods called observed-score test equating can be viewed as simple adaptations of scale-aligning methods to the problem of equating tests. This includes the linear and equipercentile methods discussed in Section 2.3.1, as well as the methods adapted to the anchor-test designs discussed in Section 2.3.2. Levine (1955) was the first application of classical test theory to the problem of equating tests, and Lord (1980) first applied IRT to test equating. Other attempts to give a theoretical foundation to test equating include Morris

(1982), Hanson (1991), and van der Linden (2000). Hanson's theorem (Holland & Dorans, 2006) is the earliest result that derives an equating function from formalizations of conditions that are related to the five equating requirements in Section 2.3.1.

Flanagan (1951) was careful to indicate the potential sensitivity of linking functions to the groups and samples used to form them. He even went so far as to state, "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. 758). This negative position is rather different from that taken later by Angoff (1971), who stated that equating relationships should be population invariant, or in his words, "…the resulting conversion should be independent of the individuals from whom the data were drawn to develop the conversion and should be freely applicable to all situations" (p. 563). Thus, both the requirement of population invariance for equating and its denial have roots that are at least 50 years old. See Kolen (2004b) for more on the history of population invariance and test equating. See also Chapters 6, 4, 12, and 10 by von Davier, Petersen, Sawyer, and Dorans and Walker, respectively, in this volume for discussions of what to do if population invariance fails to be met.

Acknowledgments. Many colleagues have helped me learn about test equating and scaling. Perhaps the most important is my colleague and co-author Neil Dorans, who has been a strong source of information and opinion for many years. In writing this chapter, I have borrowed heavily from parts of our chapter on linking and equating for the fourth edition of *Educational Measurement*. Many thanks go to Robert Brennan, Michael Kolen, Nancy Petersen, Mary Pommerich, and to my ETS colleagues, Tim Davey, Alina von Davier, Daniel Eignor, Kim Fryer, and Samuel Livingston, for their detailed reviews and comments on the material that led to this chapter.