3 Data Collection Designs and Linking Procedures

Michael J. Kolen[1]

The University of Iowa

3.1. Introduction

Scores on tests are linked using statistical procedures on data that have been collected in a systematic way. The outcome of a linking study is one or more statistically based linking functions that relate scores on one test or form to scores on another test or form. The purposes of the present chapter are to describe commonly used designs for collecting data and statistical procedures for linking scores.

The score linking situations considered are those in which scores from the tests or forms to be linked are expressed on a common metric and used for common purposes. These situations are restricted in this chapter to the linking of tests that are intended to measure the same or similar constructs. With reference to the Holland and Dorans (2006) and Holland (Chapter 2) description of types of linking method, only test form equating and concordance are considered. Predicting and scale aligning for tests measuring dissimilar constructs and vertical scaling in the Holland and Dorans (2006) and Holland (Chapter 2) framework were not considered. Vertical scaling was considered further in Patz and Yao (Chapter 14), Harris (Chapter 13), and Yen (Chapter 15). Linkages involving aggregate-level data are not addressed in this chapter. The interested reader should consult chapters by Thissen (Chapter 16), Braun and Qian (Chapter 17), and Koretz (Chapter 18).

[1] The opinions expressed in this chapter are those of the author and not necessarily of the University of Iowa.

In this chapter, the features of testing situations that influence linking are described. Equating and linking tests that are intended to measure similar constructs are distinguished. Common data collection designs and their variants for equating and for linking tests that are intended to measure similar constructs are considered. Statistical linking methods are described.

3.2. Features of Testing Situations

There have been various frameworks developed in recent years for distinguishing among and developing terminology for different types of linking (e.g., Feuer et al., 1999; Holland, Chapter 2; Holland & Dorans, 2006; Kolen & Brennan, 2004, Chapter 10; Linn, 1993; Mislevy, 1992; and the special issue of *Applied Psychological Measurement* edited by Pommerich & Dorans, 2004). The Holland and Holland and Dorans frameworks are the most up-to-date. Even these frameworks, and the associated terminology, do not emphasize important features of linking situations that are important for discussing linking designs and methods. For this reason, notation and terminology used in this chapter are in some cases different from those in typical usage.

In distinguishing among linking designs, it is important to acknowledge that the entire context of test administration affects scores on tests and can influence linking functions. For the purposes of this chapter, these features are considered in three categories: test content, conditions of measurement, and examinee population.

3.2.1. Test Content

An examinee's score on a test depends on the content of the test. Test content is considered broadly here as tasks that are presented to examinees. Standardized tests are developed with clearly defined content and statistical specifications that delineate the content areas, intended cognitive complexity, and item types to be included on a test. Features such as length of reading passages, complexity of diagrams, specifications for writing prompts, and so forth are carefully delineated in such specifications. Test specifications are an essential blueprint for test construction that provides an operational definition of test content.

3.2.2. Conditions of Measurement

Scores also depend on the conditions under which the test is administered, referred to here as *conditions of measurement*. Some of these conditions are under the control of the test developer, such as the instructions, booklet layout, answer sheet design, timing, scoring procedures, aids such as calculators, mode of administration (e.g., computer or paper-and-pencil), how items are displayed on a computer screen, and so forth. Conditions of measurement not under the direct control of the test developer include the stakes associated with test performance, the reasons an examinee is taking a test, and the type of test preparation activities.

3.2.3. Examinee Population

In aggregate, scores on tests differ for different examinee populations, such as those defined by gender, race, geographic region, or month of administration. Linking functions can differ from one examinee population to the next. Recent work has been done on examining the dependence of linking functions on examinee population. Much of this work was summarized in the special issue of the *Journal of Educational Measurement* edited by Dorans (2004a).

3.2.4. Construct Measured

The construct measured by a test clearly depends on the content of the test. The construct also depends on the conditions of measurement. For example, a test given under highly speeded administration conditions likely measures a different construct than a test given with ample time for all examinees to finish. The construct also can depend on the examinee population. For example, an English language reading comprehension test would likely measure a different construct for English language learners than for native English speakers.

3.3. Types of Linking Considered

Alternate forms of a test are built to the same test specifications. Alternate forms have nearly identical content features and differ only in the particular items that appear on the alternate forms. In operational administrations, alternate forms typically are administered under common conditions of measurement. As the term is used in the present chapter, *test*

form equating can be conducted when the test content and conditions of administration for the alternate forms to be equated are held constant. Using this restrictive definition of equating, scores on alternate forms of carefully constructed multiple-choice tests, such as the ACT® assessment multiple-choice tests, can be equated. Equating designs and methods were also considered in Cook (Chapter 5, Section 5.2), von Davier (Chapter 6, Section 6.2), Holland (Chapter 2, Section 2.4.3), and Petersen (Chapter 4).

By this definition of equating, the term *equating* is *not* appropriate for linking tests that are intended to measure similar constructs. Situations that are *not* equating include linking scores on tests that differ in content and/or conditions of measurement.

Table 3.1 provides some examples of linking situations. The upper left-hand cell of this 2 × 2 table illustrates equating, where the content and conditions of measurement are the same for the tests to be linked.

The lower right-hand cell gives situations in which both the content and conditions of measurement are not the same. This situation is typical of many in which scores on tests that are intended to measure similar constructs are linked. For example, linking scores on the mathematics test of the ACT assessment to scores on the mathematics test of the SAT ® involves tests of somewhat different content that are administered under somewhat different conditions of measurement. These sorts of linking have traditionally been referred to as concordances and they are considered in Dorans and Walker (Chapter 10), Pommerich (Chapter 11), and Sawyer (Chapter 12).

Some situations exist in which the tests differ in conditions of measurement but not in content. Examples are given in the lower left-hand portion of Table 3.1. One example is linking scores on a linear computer-based test and a paper-and-pencil test, where the same items are given in the two administration modes. This sort of situation was considered further in Eignor (Chapter 8) and Brennan (Chapter 9). There are also situations in which tests differ in content but not in conditions of measurement. Examples are given in the upper right-hand portion of Table 3.1. One example is the revision of test content specifications when there are no changes in administration conditions. This sort of situation was considered further in Liu and Walker (Chapter 7) and Brennan (Chapter 9).

All of the situations just mentioned are referred to in this chapter as examples of *linking tests intended to measure similar constructs*. In the Holland and Dorans (2006) and Holland (Chapter 2) linking categorization, the upper left-hand cell of Table 3.1 is referred to as test equating. The other three cells describe variations of what is referred to as scale aligning. In the Holland and Dorans (2006) and Holland (Chapter 2) linking categorization, equating is said to produce equivalence tables, whereas scale aligning is said to produce concordance tables.

Table 3.1. Examples of situations for linking scores on tests that differ in content and/or conditions of measurement

		Content	
		Same	Not same
Conditions of measurement	Same	Alternate forms of multiple-choice tests of the ACT Assessment	Old and new versions of a test when there has been a shift in test content, but not in administration conditions
		Alternate forms of the multiple-choice tests of the SAT	Scores for examinees who choose to take different questions on a test that allows examinee choice about which questions to answer
	Not Same	Computer-based linear and paper-and-pencil tests, when no changes are made to test content	ACT Assessment and SAT Reading achievement tests from two different publishers
		A constructed response test before and after a change in scoring rubric, assuming that the examinees are unaware of the change	Computer-adaptive and paper and-pencil tests. Tests administered in different languages

3.4. Linking Functions and Features of Testing Situations

Linking functions depend on the content of the tests, the conditions of measurement, and the population features of linking situations. The designs for data collection for linking exert control over these features of the testing situation.

Consider that scores on Test X and Test Y are to be linked. A score on Test X is represented by X and a score on Test Y is represented by Y. Linking functions depend on the content of Test X, CX, and the content of Test Y, CY. Linking functions also can depend on the population of examinees. In most situations, examinees for a linking study are sampled from an actual population, P, that differs from the target population, T, on which the linking function is ideally defined.

Linking functions also can depend on the conditions of measurement for Test X, *MX*, and Test Y, *MY*. The conditions of measurement in linking studies can differ from conditions of measurement that are considered ideal, *IX* for Test X and *IY* for Test Y.

To emphasize that linking functions can depend on all of the features of testing situations, the statistical notation for linking functions used in this chapter carries all of these important features. Consider a study in which data are collected and scores on Test X and Test Y are linked. In this study, the random-variable test score on Form X with content *CX* administered under conditions of measurement *MX* is symbolized as $X_{CX,MX}$, with particular score (realization) $x_{CX,MX}$. For Test Y with content *CY* administered under conditions of measurement *MY*, the random variable is $Y_{CY,MY}$. Using *link* for a general linking function, the notation that is used to specify a function for linking scores on Test X to scores on Test Y in a particular population, *P*, is

$$link_{Y_{CY,MY}|P}(x_{CX,MX}).$$

This function can be read as a function in population *P* for linking a score on Test X with content *CX* administered under conditions of measurement *MX* to scores on Form Y with content *CY* administered under conditions of measurement *MY*. This notation makes it clear that the linking function depends on the examinee population, the content of each test, and the conditions of measurement for Test X and Test Y.

Now also consider a situation in which the conditions of measurement are ideal and the target population, *T*, is used to define the linking function. Using similar notational conventions, this ideal linking function is specified as

$$link_{Y_{CY,IY}|T}(x_{CX,IX}).$$

Thus, this ideal linking function can differ from the actual linking function,

$$link_{Y_{CY,MY}|P}(x_{CX,MX})$$

on the population of examinees and on the conditions of measurement for Test X and Test Y.

When scores on test forms are equated, it is assumed that the content of Form X is the same as the content for Form Y, so that

$$CX = CY = C.$$

When equating, it is also assumed that the conditions of measurement for Form X and Form Y are the same, so that

$$MX = MY = M .$$

When equating using operational administrations, it is assumed that the actual conditions of measurement are ideal, so that

$$MX = MY = IX = IY = I .$$

When scores on tests that are intended to measure similar constructs are linked, it is assumed that the content of Test X and Test Y are different, so that

$$CX \neq CY .$$

In these situations, it also is assumed that the conditions of measurement for Test X and Test Y are different, so that

$$MX \neq MY .$$

When scores on test forms are equated or scores on tests are linked using special administrations or data collections, it is assumed that the actual conditions of measurement are different from the ideal conditions of measurement so that

$$MX \neq IX , \; MY \neq IY , \text{ and } \; IX \neq IY .$$

Although likely oversimplifications, these assumptions are used to highlight the importance of test content and conditions of measurement and to help compare and contrast the various designs.

3.5. Linking Designs

Commonly used designs for data collection in equating (Kolen & Brennan, 2004) are considered in this section. Counterparts of these designs for linking tests that are intended to measure similar constructs, as well as some variations, are also considered. In this section, a design is discussed first as it is used in equating and then as its counterparts and variations are used to link tests intended to measure similar constructs.

3.5.1. Random Groups Design for Equating

The random groups design for equating is diagrammed in Figure 3.1. In this design, examinees are randomly assigned Form X or Form Y. A spiraling process is often used with this design. In one method for spiraling, the alternate forms are alternated when the forms are packaged. When the booklets are handed out, the first examinee receives Form X, the second examinee receives Form Y, the third examinee receives Form X, and so on. This process leads to comparable, randomly equivalent groups taking Form X and Form Y. With the random groups equating design, the tests can be administered during standard operational administration conditions. Holland (Chapter 2, Section 2.4.3) would consider this design to be a common population design.

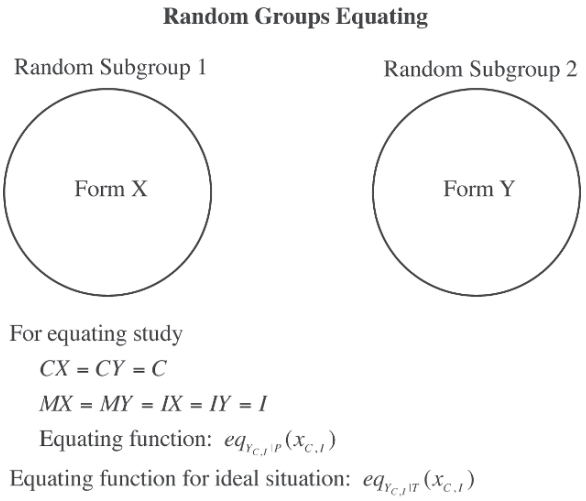

Figure 3.1. Diagram for random groups equating design.

Because this is an equating, it is assumed that the content of Form X and Form Y is the same, so $CX = CY = C$, as indicated in Figure 3.1. In an equating study using this design, the conditions of measurement for Form X and Form Y typically are identical to one another when both forms are administered in the same testing rooms under operational testing conditions. Although situations exist to the contrary, the conditions of measurement are the same for Form X and Form Y and are considered ideal when the design is implemented in an operational administration. Thus, $MX = MY = IX = IY = I$, as indicated in Figure 3.1. Using *eq* to refer to an equating function, which is a special type of linking function,

the actual equating function from the equating study is denoted $eq_{Y_{C,I}|P}(x_{C,I})$, and the ideal equating function is denoted as $eq_{Y_{C,I}|T}(x_{C,I})$, as shown in Figure 3.1. A comparison highlights that the conditions of measurement for the two forms are the same (and ideal) when equating with the random groups design. The only difference between the two equating functions is due to population. There is much evidence in the literature (see the special issue of the *Journal of Educational Measurement* edited by Dorans, 2004a) that equating functions depend little on population, so there is reason to expect that, in practice, the actual and ideal equating functions will be very similar.

In the random groups equating design, the difference between group-level performance on the two forms is taken as a direct indication of the difference in difficulty for the two forms. Various statistical procedures, which require only minimal statistical assumptions, are available to estimate equating functions that equate scores on Form X and Form Y.

3.5.2. Random Groups Design and Variations for Linking

A random groups design can be implemented for linking tests that are intended to measure similar constructs. This design is illustrated in Figure 3.2. One way that Figure 3.2 differs from Figure 3.1 is that *test* replaces *form*. To apply this design to linking, examinees are randomly assigned to be administered Test X and Test Y. Holland (Chapter 2, Section 2.4.3) would consider this design to be a common population design.

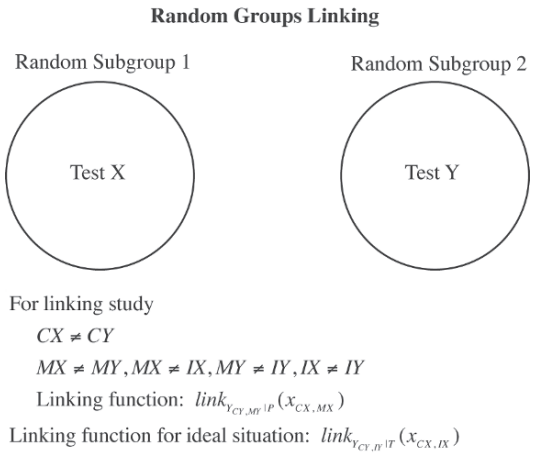

Figure 3.2. Diagram for random groups linking design.

Compared to random groups equating, the random assignment can be much more difficult to implement when the conditions of measurement for Test X and Test Y differ. For example, if the time limits for Test X and Test Y differ, it would be difficult to administer both tests in the same room. As another example, it would be difficult to administer a computer-based test and a paper-and-pencil test in the same room. In these linking situations, examinees could be assigned to take Test X or Test Y ahead of time. Students assigned to Test X would take the test in one room and students assigned to Test Y would take the test in another room.

Given these administration complications, Test X and Test Y, in general, cannot be administered in a standard operational administration when using this design. In this case, a special linking administration is needed. If the conditions of measurement in the linking study differ from those used operationally, then the conditions of measurement in the linking study likely differ from the ideal conditions of measurement. In addition, the examinees included in the linking study, out of necessity, might not be representative of the target population of examinees.

For the linking design illustrated in Figure 3.2, it is assumed that Test X and Test Y differ in content, so $CX \neq CY$. In addition, the conditions of measurement for Test X and Test Y differ from one another because each test is different and each is administered under its own conditions of measurement. Because the linking typically requires a special data collection, the conditions of measurement likely differ from ideal conditions of measurement. Thus, as indicated in Figure 3.2, in general, $MX \neq MY$, $MX \neq IX$, $MY \neq IY$, and $IX \neq IY$. The linking function from the linking study, $link_{Y_{CY,MY}|P}(x_{CX,MX})$, can differ from the ideal linking function, $link_{Y_{CY,IY}|T}(x_{CX,IX})$, due to differences in content, differences in conditions of measurement for the tests, and differences in population. When Test X and Test Y differ in content, there is evidence in the literature to suggest that the linking relationship will depend on the population (see the special issue of the *Journal of Educational Measurement* edited by Dorans, 2004a).

To avoid the problems of having to assign students within a school to take different tests, a variation of this design is sometimes used where random assignment is conducted at the school level. This design is referred to as the *random groups design—randomization by school*. In this variation, a list of schools to be included in the study is constructed and the schools are randomly assigned to take either Test X or Test Y. Note that the unit of randomization is the school. To achieve reasonable linking precision, the number of students that must be tested is, in general, too large to be practicable.

3.5.3. Single Group Design with Counterbalancing for Equating

The single group design with counterbalancing for equating is illustrated in Figure 3.3. In this design, each examinee takes Form X and Form Y, in counterbalanced order. Counterbalancing is needed because examinee performance can differ depending on whether a form is taken first or second, due to such factors as practice and fatigue. One randomly chosen subgroup of examinees is administered Form X first. A second randomly chosen subgroup is administered Form Y first. Holland (Chapter 2, Section 2.4.3) would consider this design to be a common population design.

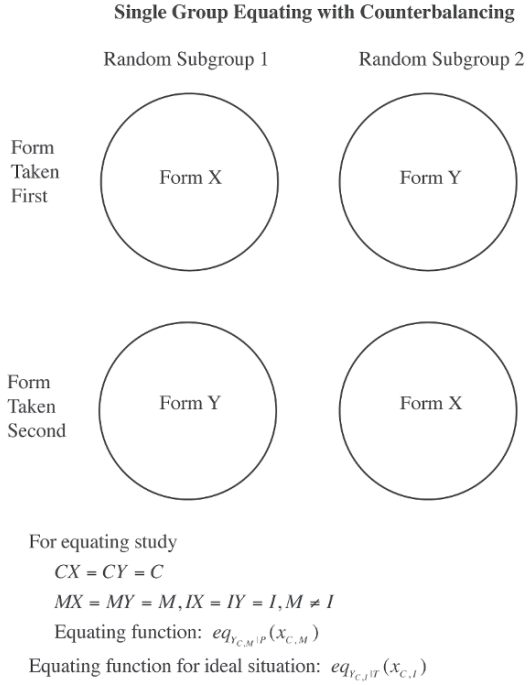

Figure 3.3. Diagram for single group with counterbalanced equating design.

A special study is required when using this design, because examinees normally do not take two test forms in operational administrations. One way to administer the forms in this design is to construct test booklets that contain both forms. Half of the booklets contain Form X followed by Form Y. The other half of the booklets contains Form Y followed by Form X. The booklets are packaged in a spiraled manner and distributed in such a way that the first examinee in a room is administered Form X first followed by Form Y, the second examinee is administered Form Y

followed by Form X, and so forth. The first and second forms are administered under separate time limits.

Refer again to Figure 3.3. The portion of the design labeled *form taken first* is identical to the random groups design shown in Figure 3.1. Thus, equating could be conducted using only the form taken first. To take full advantage of this design, data from the *form taken second* are used. However, the form taken second is administered under atypical conditions of measurement. In practice, examinees do not take two forms of a test. Thus, the data on the test taken second can be used only if the equating relationship for the form taken second can be shown to be the same as the equating relationship for the form taken first. If these equating relationships differ, then a *differential order effect* is said to occur. If this effect is substantial, then the data on the test administered second might need to be disregarded.

When alternate forms of a test are equated, there is little reason to expect that differential order effects occur because the content of the two forms is the same and the only difference in conditions of measurement is test order. When a differential order effect does not exist, the data from the two orders can be pooled. In this case, each examinee has scores on two forms, and serves as his or her own control. Consequently, for a particular sample size, this design leads to much more precise estimates of equating relationships than does the random groups design.

The single group design with counterbalancing is administered in a special study, which can lead to the conditions of measurement for this design being different from those for an operational administration. These different conditions of measurement can lead to differences between the equating function estimated in this design and the ideal equating function.

When equating with this design, it is assumed that the content of the two forms is the same, so $CX = CY = C$, as indicated in Figure 3.3. Assume that there is no differential order effect, so that the conditions of measurement for Form X and Form Y are considered the same. Thus, as indicated in Figure 3.3, $MX = MY = M$, where M represents the conditions of measurement in the study. In the ideal situation, $IX = IY = I$, where I represents the ideal conditions of measurement. Because a special study is used, the conditions of measurement for the study likely are different from the ideal conditions of measurement. Thus, in general, with this design, $M \neq I$. In this situation, as indicated in Figure 3.3, the equating function for an equating study is denoted as $eq_{Y_{C,M}|P}(x_{C,M})$ and the ideal equating function is denoted as $eq_{Y_{C,I}|T}(x_{C,I})$. This notation illustrates that the equating function for the equating study differs from that for the ideal equating function due to differences in conditions of measurement and differences in examinee population.

In some situations, what Holland (Chapter 2, Section 2.4.3), Holland and Dorans (2006), and Kolen and Brennan (2004) referred to as a single group design might be considered. In the single group design, examinees are administered the two tests to be equated, but the order of administration is not counterbalanced. The portion of Figure 3.3 for random subgroup 1 is an example of this design, where all of the examinees take Form X followed by Form Y. When order effects exist, there is no way to estimate their magnitude or to adjust the equating relationship for the effect of order when using the single group design. Thus, it is difficult to justify the use of the single group design in practical equating contexts.

3.5.4. Single Group Design with Counterbalancing and Variations for Linking

The single group design with counterbalancing for linking is illustrated in Figure 3.4. One way that Figure 3.4 differs from Figure 3.3 is that *test* replaces *form*. In this linking design, the content of the two tests is assumed to differ, so $CX \neq CY$, as indicated in the figure. This design can be particularly difficult to administer when linking two tests that are intended to measure similar constructs. Typically, in this situation the conditions of measurement are different for the two tests (i.e., $MX \neq MY$), so it is not possible to administer both tests in the same room. Holland (Chapter 2) would consider this design to be a common population design.

For example, suppose that Test X is a paper-and-pencil test and Test Y is a computer-based test. It likely would not be feasible to administer both modes in the same testing room at the same time. Instead, examinees are assigned to condition ahead of time, and special procedures are used for when and how the examinee takes each of the assigned tests in the order required by the design.

Proper administration of this design requires that examinees be randomly assigned to condition (*test taken first*) and that the tests be administered appropriately. In addition, it is necessary to assess whether differential order effects occur. It seems much more likely that differential order effects will be present when linking tests that are intended to measure similar constructs than when equating test forms, because the conditions of measurement for the two tests differ. For example, the effect of first taking a computer-based test on subsequent scores on a paper-and-pencil test likely differs from the effect of first taking a paper-and-pencil test on subsequent scores on a computer-based test. If so, then a differential order effect occurs, and the data for the test taken second might

need to be disregarded. However, disregarding data from the test administered second leads to a serious loss in linking precision.

As indicated near the bottom of Figure 3.4, when linking Test X to Test Y using the single group design with counterbalancing for linking and its variations, test content differs, the conditions of measurement differ for Test X and Test Y, and these conditions of measurement differ from the ideal conditions of measurement. Also, as indicated at the bottom of Figure 3.4, the linking function from the study, $link_{Y_{CY,MY}|P}(x_{CX,MX})$, differs from the ideal linking function, $link_{CY,IY|T}(x_{CX,IX})$, due to differences in content, differences in conditions of measurement for the tests, and differences in examinee population.

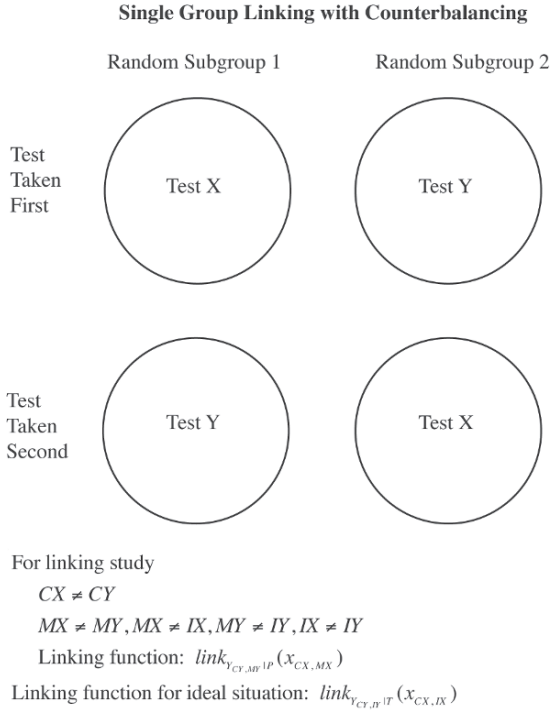

Figure 3.4. Diagram for single group with counterbalancing linking design.

Because of the serious practical difficulties in administering the single group design with counterbalancing in many linking situations, variations of this design often are used in practice. In one variation, the random assignment to condition is done by school. This design is referred to here

as the *single group design with counterbalancing for linking—randomization by school*. For example, using a random selection procedure, one set of schools is assigned to be administered Test X first and a second set of schools is assigned to be administered Test Y first. In this case, school is the unit of randomization, which leads to substantial loss of precision when assessing whether there is a differential order effect. If a differential order effect cannot be ruled out, then a linking function calculated by pooling data would not necessarily control for differences in conditions of measurement for the ideal as compared to the actual linking functions.

Another variation of this design is one in which examinees are found who have taken both of the tests to be linked, with examinees found who have taken the tests in both orders. This design is referred to here as the *single group design with counterbalancing for linking—naturally occurring groups*. This sort of design is used, for example, to link scores on the ACT assessment to scores on the SAT exam. Pommerich (Chapter 11), Dorans and Walker (Chapter 10), and Sawyer (Chapter 12) considered situations in which this design is used. In this design, some examinees are found who have taken one test first and other examinees are found who have taken the other test first. The time between administrations can vary, as can the test forms. In addition, the population of examinees who take the two tests can differ considerably from the general population of test-takers. In this design variation, differences in conditions of measurement as compared to ideal conditions can differ widely and are, for the most part, uncontrolled.

The single group design, where all of the examinees take the tests in the same order, also might be considered for use in linking. If this design is used, the linking function will be affected by order effects by an unknown amount, making it difficult to justify the use of the single group design for linking.

3.5.5. Common-Item Nonequivalent Groups Design for Equating

The common-item nonequivalent groups design for equating is illustrated in Figure 3.5. This design is used when only one form can be administered per test date. In this design, Form X and Form Y have a set of items in common. Examinee Group 1 takes Form X and examinee Group 2 takes Form Y. The two groups of examinees might test on different test dates. With this design, examinee Group 1 is considered to differ systematically from examinee Group 2. This design was referred to as the nonequivalent groups anchor test (NEAT) design by Holland (Chapter 2, Section 2.4.3).

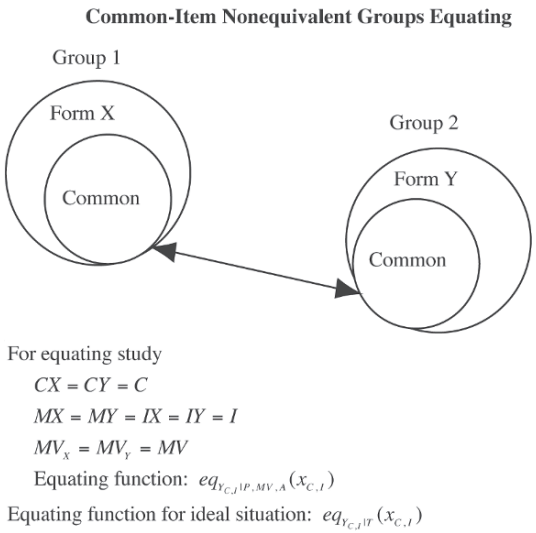

Figure 3.5. Diagram for common-item nonequivalent groups equating design.

This design has two variations. When the score on the set of common items contributes to the examinee's score on the test, the set of common items is referred to as internal. Typically, these items are interspersed among other scored items. When the score on the set of common items does not contribute to the examinee's score, the set of items is referred to as external. Typically, external common items are administered in a separately timed section.

Scores on the common items provide direct information on how the performance of examinee Group 1 differs from the performance of examinee Group 2. The set of common items is chosen to proportionally represent the total test forms in content and statistical characteristics. To ensure that the common items behave the same way on the two forms, each of the common items is identical on the two forms and is in a similar position in the test booklet.

When conducting equating using this design, strong statistical assumptions are required to disentangle form differences from examinee group differences. Especially when there are large group differences, the set of assumptions chosen can have a substantial effect on the equating results.

Because this is an equating study, the content of Test X and Test Y are the same (i.e., $CX = CY = C$) as shown in Figure 3.5. The measurement conditions for Form X and Form Y often can be considered to be the same and ideal when this design is conducted in operational administration so that $IX = IY = MX = MY = I$, as indicated in Figure 3.5.

The actual equating relationship depends on the set of common items. Let V represent score on the common items, let MV_X represent the conditions of measurement for the common items as administered with Form X, and let MV_Y represent the conditions of measurement for the common items as administered with Form Y. Assume that the context of the common items is the same for Form X and Form Y and that the common items accurately reflect the content of the total scores. In this case, it seems reasonable to assume that the conditions of measurement are the same for the common items, regardless of test form. Denoting the common conditions of measurement as MV ($MV_X = MV_Y = MV$). The actual equating relationship also depends on the set of assumptions that are made, denoted as A.

Notation for the equating function is expressed in Figure 3.5 as $eq_{Y_{C,I}|P,MV,A}(x_{C,I})$. The ideal equating function does not depend on the common items, because it is a relationship between scores on Form X and Form Y. So, the ideal equating function is expressed as $eq_{Y_{C,I}|T}(x_{C,I})$ in Figure 3.5. Comparing these two functions highlights that the conditions of measurement for the two forms are the same (and ideal) when equating with this design using operational administrations. The differences between the two equating functions are due to differences in population and the statistical assumptions used to estimate the equating function.

3.5.6. Anchor-Test Nonequivalent Groups Design for Linking

The *anchor-test nonequivalent groups design* illustrated in Figure 3.6, used to link tests that are intended to measure similar constructs, has similarities to the common-item nonequivalent groups design. In this design, Test X is administered to one group, Test Y is administered to a second group, and an anchor test, Test V, is administered to both groups. A major requirement in the common-item nonequivalent groups design for equating is that the content of the common items adequately represents the content of Form X and Form Y. When the content of Test X and Test Y differ, it is impossible for the common items to adequately represent the content of both Tests X and Y. Thus, the common-item nonequivalent groups design cannot be used when linking tests that are intended to measure similar constructs. Instead, the anchor-test nonequivalent groups design, which does not require that the anchor test have the same content as Test X and Test Y, is used. Linking using this design would fall under the category *concordance* using an anchor measure in the framework presented by Holland (Chapter 2).

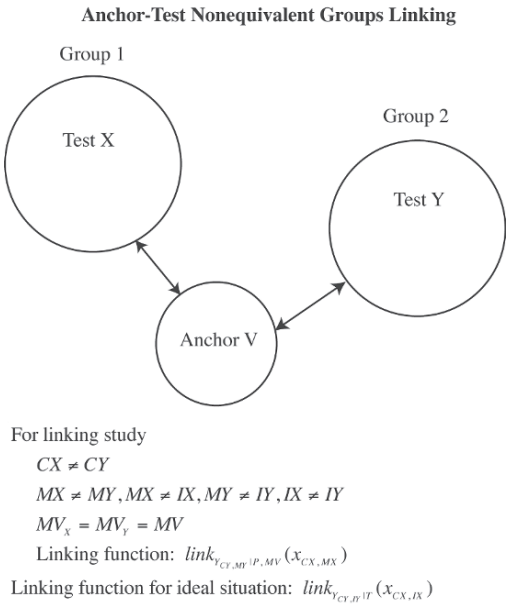

Figure 3.6. Diagram for anchor-test nonequivalent groups linking design.

In the anchor-test nonequivalent groups design, it is crucial that the conditions of measurement for the anchor test are the same for the group taking Test X$\left(MV_X\right)$ and Test Y $\left(MV_Y\right)$. Otherwise, examinee group differences are completely confounded with differences in conditions of measurement for the two groups. So, in Figure 3.6, $MV_X = MV_Y = MV$.

In linking using this design, the conditions of measurement for Test X and Test Y typically differ from one another. In these studies, the conditions of measurement for Test X and Test Y also could differ from ideal conditions of measurement. For this reason, the actual linking function in Figure 3.6 is $link_{Y_{CY,MY}|P,MV,A}\left(x_{CX,MX}\right)$. The ideal linking function in Figure 3.6 is $link_{Y_{CY,IY}|T}\left(x_{CX,IX}\right)$, which makes explicit that the ideal conditions of measurement for Test X can differ from the ideal conditions of measurement for Test Y. By comparing these functions, it can be seen that the actual function can differ from the ideal function due to differences between the actual and the ideal conditions of measurement for Test X, differences between the actual and the ideal conditions of measurement for Test Y, and differences in population. The assumptions

(*A*) can also contribute to differences between these two functions. As is made clear in the discussion of statistical methods later in this chapter, it is unlikely that the statistical assumptions made in this linking design hold in situations where Test X and Test Y differ in content and the group of examinees taking Test X differs substantially from the group of examinees taking Test Y.

3.6. Linking Procedures

In this section, statistical procedures for equating alternate forms and linking scores on tests intended to measure similar constructs are considered. Equating and linking methods were described in detail elsewhere (e.g., Holland & Dorans, 2006; Kolen & Brennan, 2004), so only an overview is provided here.

As described earlier, the score linking situations considered were those in which scores from the tests or forms to be linked are expressed on a common metric and used for a common purpose. To address these situations, only *symmetric* statistical linking functions were considered (see Holland, Chapter 2).

In this section, overviews of traditional and item response theory (IRT) methods for equating are presented. Then the application of some the methods to linking tests that measure similar constructs is considered.

3.6.1. Traditional Statistical Methods for Equating

The intent of traditional methods of equating is for scores on alternate forms to have the same score distributional characteristics in a population of examinees, after the scores are transformed to a common scale. *Mean equating* results in scores having the same mean on the common scale. Using a linear transformation, *linear equating* results in scores having the same mean and standard deviation on the common scale. Using a nonlinear transformation, *equipercentile equating* results in scores on alternate forms having approximately the same score distribution on the common scale. Focus in this section is on equipercentile methods.

Equipercentile equating functions are defined for a population and for tests given under particular conditions of measurement. Define F_T as the cumulative distribution of scores on Form X in population T, G_T as the cumulative distribution of scores on Form X in population T, G_T^{-1} as the inverse of G_T, and $x_{C,I}$ and $Y_{C,I}$ as defined earlier. Based on results

presented by Braun and Holland (1982), when scores are continuous, Form X and Form Y measure content *C,* and the forms are administered under ideal conditions of measurement *I,* the equipercentile equating function for population *T* can be expressed as

$$eq_{Y_{C,I}|T}\left(x_{C,I}\right) = G_T^{-1}\left[F_T\left(x_{C,I}\right)\right]. \tag{3.1}$$

By substituting different subscripts in Equation 3.1, the function can be defined for other populations or for other conditions of measurement. For example, the equipercentile equating function for forms administered under other than ideal conditions of measurement, *M,* to examinees from population *P* is expressed as

$$eq_{Y_{C,M}|P}\left(x_{C,M}\right) = G_P^{-1}\left[F_P\left(x_{C,M}\right)\right]. \tag{3.2}$$

Estimates of the cumulative distribution functions can be used with Equations 3.1 and 3.2 to produce an estimated equating function.

Because scores on tests typically are discrete, a procedure is used to *continuize* scores so that the equations can be applied. Traditionally, percentiles and percentile ranks are used to continuize scores. If scores are integers, percentiles and percentile ranks can be thought of as continuizing scores by uniformly spreading the score density at an integer score over the range $x-.5$ to $x+.5$. von Davier, Holland, and Thayer (2003) provided an alternate scheme for continuizing scores referred to as the *kernel method*. Using the kernel method, the score density at an integer score is spread using a Normal distribution. Either of these approaches leads to continuous scores that can be equated using Equations 3.1 and 3.2.

Smoothing methods often are used with estimates of equipercentile equating functions to reduce sampling error. In *presmoothing*, the score distributions are smoothed. The *log-linear smoothing method*, which is summarized by Kolen and Brennan (2004) and by von Davier et al. (2003), is an often-used presmoothing method. In *postsmoothing*, the equipercentile function is smoothed directly. The *cubic spline postsmoothing method* described by Kolen and Brennan is an often-used postsmoothing method.

3.6.1.1. Random Groups and Single Group with Counterbalancing Designs

After data are collected using the random groups design, equipercentile equating, continuization, and smoothing procedures are applied. For the single group design with counterbalancing, after deciding on whether data from the forms taken second can be used, similar procedures are followed.

3.6.1.2. Common-Item Nonequivalent Groups Design

Traditional equating methods using the common-item nonequivalent groups design (referred to as the NEAT design by Holland, Chapter 2) are more complicated. In this design, statistical assumptions are required to disentangle form and group differences.

In one class of methods, sometimes referred to as *poststratification methods*, the following nontestable assumptions are made: the regression of X on V is the same in examinee Group 1 and Group 2 and the regression of Y on V is the same in Group 1 and Group 2. In the *Tucker linear method*, assumptions are made regarding linear regressions. In the *frequency estimation equipercentile method*, assumptions are made regarding nonlinear regressions. A synthetic population is defined as a combination of the populations from which Group 1 and Group 2 are sampled. The equating function is based on this population. The assumptions made in poststratification methods seem less likely to hold when Group 1 and Group 2 differ substantially in proficiency.

Smoothing methods can be applied when conducting the frequency estimation equipercentile method. von Davier et al. (2003) summarized a log-linear smoothing in the context of the kernel method. Kolen and Brennan (2004) summarized a cubic spline postsmoothing method in which a cubic spline function is fit to the unsmoothed equipercentile equivalents.

In another class of methods for linear equating, referred to as *Levine methods*, an assumption is made that true scores on X and V in Group 1 are perfectly linearly correlated and that true scores on Y and V in Group 1 are perfectly linearly correlated. This assumption seems less likely to hold when the common items measure a construct that differs from the construct measured by the alternate forms.

A third class of traditional methods for the common-item nonequivalent groups design are *chained methods*. In these methods, X is linked to V in Group 1, V is linked to Y in Group 2, and these two linkings are chained together. A *chained linear method* and a *chained equipercentile method* have been developed.

3.6.2. IRT Statistical Methods for Equating

Unidimensional IRT models assume that examinee proficiency can be described by a single latent variable, θ, and that items can be described by a set of parameters or curves that relate proficiency to probability of correctly answering the item (Lord, 1980). Unidimensional IRT models have been developed for use with test items that are dichotomously scored or polytomously scored. IRT models are based on strong statistical assumptions. The θ-scale has an indeterminate location and spread. For this reason, one θ-scale sometimes needs to be converted to another linearly related θ-scale. If summed scores are to be used, there are two steps in IRT equating (Kolen & Brennan, 2004). First, the θ-scales for the two forms are considered to be equal or are set equal. Then summed score equivalents on the two forms are found.

In many situations, the parameter estimates for the two forms are on the same θ-scale without further transformation. The typical situation in which a transformation of the θ-scale is required is in the common-item nonequivalent groups design when Form X and Form Y parameters are estimated separately.

After the parameter estimates are on the same scale, *IRT true-score* and *IRT observed-score* methods can be used to relate summed scores on Form X to summed scores on Form Y. In IRT true-score equating, the true-score on one form associated with a given θ is considered to be equivalent to the true score on another form associated with that same θ.

Item response theory observed-score equating uses the item parameters estimated for each form along with the estimated distribution of ability for the population of examinees to estimate the distributions of summed scores for Form X and Form Y. Standard equipercentile equating procedures are used to equate these two smoothed distributions. As Holland and Dorans (2006) noted, IRT observed-score equating can be viewed as an equipercentile equating of presmoothed score distributions that are consistent with the assumptions of an item-level response model.

Any application of unidimensional IRT models requires that all of the items measure the same unidimensional proficiency, that the item responses are conditionally independent, and that the relationship between proficiency and probability of correct response follows the particular IRT model used.

3.6.3. Methods for Linking Tests Intended to Measure Similar Constructs

Tests intended to measure similar constructs often are linked using the same statistical methods used for equating. However, certain complications need to be addressed.

In some circumstances, when using equipercentile methods, presmoothing methods can be difficult to apply because the distributions might be expected to be irregular. For example, in linking scores on the ACT and SAT, integer-scale scores on the two tests are linked. For some test forms, the use of integer-scale scores can cause certain scale scores to be reported more often than adjacent scale scores because of the way the conversion to integers happens to be applied. In these situations, the scale score distribution is expected to be irregular. Such expected irregularities can lead to complications with presmoothing methods. For this reason, Kolen and Brennan (2004) used postsmoothing methods to link scale scores from different tests.

Item response theory methods can be used only in those situations in which the tests that are linked can be considered to measure the same proficiency and in situations in which item-level response data are available. For example, IRT methods would not be used to link ACT and SAT scores, because the tests do not measure the same proficiency and item-level data are typically unavailable when the tests are linked.

The statistical procedures for linking scores on tests intended to measure similar constructs with the anchor-test nonequivalent groups design (referred to as the NEAT design by Holland, Chapter 2) often are the same statistical procedures as those for equating alternate forms with the common-item nonequivalent groups design. In applying these procedures, it is important that the anchor test be administered under the same conditions of measurement for the two tests, otherwise the linking results will be misleading. For example, consider linking a paper-and-pencil to a computer-based test using the anchor-test nonequivalent groups design. Suppose that the examinees taking the computer-based test take the anchor test on the computer and that the examinees taking the paper-and-pencil test take the anchor test under paper-and-pencil conditions. In this case, group differences are completely confounded with mode of administration effects, and it is impossible to use data collected to disentangle these effects. To disentangle these effects, it would be necessary to administer the same anchor test to both groups under the same conditions of measurement. For example, a paper-and-pencil anchor test might be administered to both groups.

When using the anchor-test nonequivalent groups design, it is important to consider the effects of violations of statistical assumptions. Recall that

poststratification methods require that regressions of *X* on *V* and *Y* on *V* be the same for the groups taking Test X and Test Y. The chained methods require an assumption of population invariance of the links between Test X and anchor Test V and between anchor Test V and Test Y. These assumptions are less likely to hold as the extent of the differences in content or administration conditions for Test X and Test Y increase and to the extent that the differences in the proficiencies of the group taking Test X and Test Y increase. When using IRT methods with this design, an assumption is made that all items on Test X, Test Y, and the anchor test measure the same proficiency. This assumption is unlikely to hold for most situations in which scores on tests that measure similar constructs are linked.

When using the anchor-test nonequivalent groups design for linking scores on tests of different content, the anchor test cannot adequately represent the content of both Test X and Test Y. In this case, the linking results likely depend on the particular anchor chosen. If possible, the linking can be conducted using different anchor tests and the sensitivity of the linking to choice of anchor test assessed. In addition, the standard methods might be modified to accommodate the use of multiple anchors in a single linking.

3.7. Summary and Conclusions

Notation and terminology were used in this chapter to distinguish among designs, linking functions, and linking results. The notation incorporated population, conditions of measurement, and content. This notation makes explicit those factors on which linking functions depend. Terminology used with equating designs was expanded from typical terminology to distinguish between designs used in linking and equating. For example, the use of the term *common-item nonequivalent groups design* for equating and the term *anchor-test nonequivalent groups design* for linking tests that measure similar constructs serves to highlight the substantial differences between these designs (Holland, Chapter 2, referred to both of these designs as the NEAT design). In particular, in equating, the content of the set of common items represents the content of Form X and Form Y, whereas when linking tests intended to measure similar constructs, the content of the anchor test typically does not represent the content of both Test X and Test Y. Further developments in notation and terminology should serve to better distinguish among different linking situations, to display important differences among the designs, and to highlight the effects of factors such as content, conditions of measurement, and population on linking results.

When conducting equating, Form X and Form Y have the same content and typically are administered under the same conditions of measurement, providing significant statistical control. Equating can be expected to provide reasonable results, and the statistical assumptions required for conducting equating can be expected to hold reasonably well in a variety of situations.

When linking scores on tests that are intended to measure similar constructs, Test X and Test Y typically have somewhat different content and are administered under different conditions of measurement to examinees from populations that differ from the target population. Thus, there is significantly less statistical control exerted in these situations than in equating situations. In addition, data collection designs often are very difficult to implement properly and statistical assumptions often are violated. Because of these complications, linking of scores on tests that measure similar constructs likely depends on the examinee population and on the conditions of measurement.

Because of these dependencies, the sensitivity of linking functions to variations in conditions of measurement and population should be assessed. If there is substantial variation, then either reporting different linking relationships for different conditions of measurement and populations or not reporting the relationships should be strongly considered. In any case, when presenting the results of linking, test content, conditions of measurement, and population should be clearly specified.

Acknowledgments. The author thanks Robert L. Brennan, Neil J. Dorans, and Mary Pommerich for their detailed reviews and comments on earlier versions of this chapter.