8 Linking Scores Derived Under Different Modes of Test Administration

Daniel R. Eignor[1]

Educational Testing Service

8.1. Introduction

All established testing programs that develop computer-based versions of paper-and-pencil tests, particularly computer-adaptive tests (CATs), typically need to link the scores derived from the two administration modes. Linking is necessary because computer-based and paper-based testing will likely occur together, at least for some transition period. Further, even if paper-and-pencil testing can be immediately phased out when the computer-based test (CBT) is introduced, scores from the computer-based version will, in many cases, need to be reported on the scale that existed for the paper-and-pencil test until such time that paper-based scores are no longer accepted.

All of the above considerations necessitate that a linking study between scores from the two modes of test administration be conducted. Typically, the scores from the newer computer-based mode of administration will be linked to scores from the paper-and-pencil mode of administration and the scores from the two administrations will be reported as if they were interchangeable. The degree to which the linked computer-based and paper-and-pencil scores can be treated as interchangeable will depend on a number of different factors, the most important being the nature of the computer-based test itself.

The purposes of this chapter are twofold: (a) to clarify when a linking of scores between a computer-based test and a paper-and-pencil test can be considered to result in scores that are interchangeable and (b) using

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

available literature on the topic to provide descriptions of the ways one might design a linking study to relate scores on computer-based and paper-and-pencil-based tests.

8.2. Background

Holland and Dorans (2006) developed general definitions for a wide variety of linking methods, three of which will be important in the context of relating scores on computer-based and paper-and-pencil tests: equating, calibration, and concordance. (See Holland, Chapter 2, for a detailed discussion of types of linking.) Equating is used to refer to linking between two test forms that measure the same construct at the same level of difficulty and with the same level of reliability. Equated scores can be treated as being truly interchangeable. Calibration is used to refer to linking between two test forms that measure the same construct at approximately the same level of difficulty, but with different levels of reliability. Calibrated scores are typically treated as though they were interchangeable, although there are real questions as to whether this is appropriate. The use of a common reported score scale with scores from tests that have been calibrated actually encourages score users to use the results as if they came from an equating because there will be nothing about the nature of the scale that will help users understand that a calibration and not an equating has been done. Concordance is used to refer to linking between two different tests that measure similar constructs with somewhat similar levels of difficulty and reliability. Scores that have been concorded cannot be treated as being interchangeable.

In addition to the above definitions, it will be useful to employ two additional terms that are similar to those used by Hanson, Harris, Pommerich, Sconing, and Yi (2001): sets of equivalent scores and sets of scores that are equivalent in appearance only. A weak definition of sets of equivalent scores is that the two sets share the same raw score mean. A stronger definition is that the two sets share the same raw score mean, variance, and distribution of scores. Sets of scores that are identical in appearance only share the same raw score mean or, in the stricter sense, the same raw score mean, variance, and distribution of scores, but the scores themselves do not convey the same meaning. The scales for the two tests have been aligned, but the nature of the scores has not changed. As an example, two sets of scores might have the same raw score mean, variance, and distribution, but two scores that appear to be the same might not measure with the same level of precision; that is, the two scores might not have the same conditional standard error of measurement.

The term “linear CBT” is often used to describe a paper-and-pencil form that is administered via computer. All that differs is the mode of administration. Linkings of scores from paper-and-pencil and linear CBT modes are expected to be equatings. Whether the scores can be appropriately treated as interchangeable is an empirical question. In other instances when CATs are created, the assumption of strict interchangeability of scores from the two modes will be less appropriate. When linking scores on CATs and paper-and-pencil tests, the relationship between the scores from the two modes can at best be characterized as a calibration. Finally, there are instances in which the new computer-based test has been purposely constructed to differ from the paper-and-pencil test, either through the employment of innovative item types or through the updating of test content. In this situation, scores from the two modes cannot be considered to be interchangeable, although score users might want cut-scores on the computer-based test to be aligned with cut-scores on the paper-and-pencil test. In such a case, a concordance relationship could be established between scores across the two modes of administration. In sum, depending on the relationship between scores on paper-and-pencil and computer-based versions of the test, a linking between scores from the two modes can potentially be considered as an equating of the scores, a calibration of the scores, or a concordance between the scores. It should be noted that linear and curvilinear linking procedures, typically applied in the equating context, can also be used to calibrate scores or to bring about a concordance between scores. Most often, a curvilinear procedure, such as equipercentile linking, is employed in these contexts.

Regardless of the actual form of the linking between the scores from the two modes of administration, a data collection design must be employed to collect data to conduct the linking. (See Kolen, Chapter 3, for a description of data collection designs and a discussion of the importance of measurement conditions to linking.) Data collection designs for linking tests that are described in the literature (see also Angoff, 1984) were developed for parallel or close to parallel forms of examinations given via the same administration mode, most typically the paper-and-pencil mode. Applications of these designs to link scores derived from different modes of administration have, at times, provided results that are questionable. Questionable linkages have particularly occurred when one score is derived from a CAT and the other score is derived from a paper-and-pencil administration. As a result, variations on the standard designs in Angoff (1984) have sometimes been employed. For instance, straightforward implementation of the single group counterbalanced design in which the computer-based and paper-and-pencil tests are given contiguously in the same testing session has often produced linking results of a questionable nature. By administering the tests in a noncontiguous fashion, acceptable linking results have been produced.

In the sections that follow, a number of topics relevant to linking scores across different modes of administration will be discussed. In the next section, issues that cause scores from paper-and-pencil and certain computer-based versions of tests to lack the level of comparability brought about by an equating will be discussed. The focus will be on CATs and paper-and-pencil tests. The following section will discuss implementations of data collection designs in the context of linking scores derived from different modes of administration. Linking studies of this nature that have been documented in the literature will be discussed. Finally, the last section of the chapter will provide a summary and reflections on the linking of scores derived from different modes of test administration.

8.3. Comparability Issues Involving Scores from Computer-Based and Paper-and-Pencil Tests

The focus of this section will be on issues that cause linked scores on CATs and paper-and-pencil tests not to be comparable at the level brought about by an equating. First, however, it is useful to talk about comparability issues in the context of linking scores on linear CBTs and paper-and-pencil tests. When the same form is administered in the linear CBT and paper-and-pencil modes, the only thing that can keep scores from being equivalent across modes is the manner in which the items are presented on screen. Mazzeo and Harvey (1988) discussed many of the item presentation issues that might cause differences between scores on linear CBTs and paper-and pencil tests utilizing the same form. An updated discussion of item presentation issues is found in Pommerich (2004). Probably foremost among the issues here is how to present reading passages and items on screen. With linear CBT and paper-and-pencil versions of different test forms, scores can be affected by differences in difficulty caused by the different modes of presentation and by differences in difficulty across items. Both can keep the two sets of raw scores from being equivalent. However, the resulting scores in most instances can be linked and reported on the same reported score scale.

Score comparability issues are not a concern when forming concordances between scores on tests, be they scores on linear CBT and paper-and-pencil tests or scores on CAT and paper-and-pencil tests, because the tests in question are typically not designed to yield comparable scores. As pointed out, however, in the chapters on concordance in this volume, not all concordances are of equal quality. Typically, there will be no attempt made to report scores on a common scale. Care must be taken, however, to ensure that score users do not treat the related scores as though they were equivalent (i.e., to substitute the paper-based aligned score for a particular computer-based score in the score reporting process). Although the scales may be lined up, the scores do not mean the same thing.

Comparability issues are of concern when linking scores on CATs and paper-and-pencil tests, and a number of recent articles in the literature have focused on this issue. Perhaps the most comprehensive treatment of these issues can be found in Wang and Kolen (2001). Kolen and Brennan (2004) provided a somewhat briefer but still thorough treatment of the issues, some of which are discussed below. If these issues did not exist, it might be possible to consider a linking between scores on a CAT and a paper-and-pencil test to constitute an equating. Much of the earlier literature on the topic of linking these sorts of scores did consider the linking to constitute an equating; see Eignor (1993), McBride, Corpe, and Wing (1987), and Segall (1995).

Perhaps the first issue that comes to mind is not really an issue at all. It has to do with whether the scores from CATs and paper-and-pencil tests can be considered comparable because of possible content differences. Modern item selection algorithms used with CAT, such as those described by Stocking and Swanson (1993) or van der Linden and Reese (1998), ensure that the content coverage across the two tests is comparable, although the comparability will likely be proportional in nature, given that CATs are typically shorter than paper-and-pencil exams. Eignor, Stocking, Way, and Steffen (1993) discussed how content specifications can be treated in CATs when using the Stocking and Swanson approach so as to have the content parallel that of a paper-and-pencil test.

Test administration conditions that differ between the CAT and paper-and-pencil form might contribute to a lack of comparability. CATs are given under conditions where examinees must respond to the current item before they can receive the next item, and they are not allowed to go back and review or change items to which they have previously responded. In paper-and-pencil format, examinees can skip items and can go back and review or change previously provided responses to items.

The manner in which CAT and paper-and-pencil tests are scored can also contribute to a lack of comparability. Whereas paper-and-pencil tests are typically either number-right or formula scored, with CATs the final score is an item response theory (IRT)-based ability estimate. Typically, the ability estimate is based on a sum of weighted item responses, whereas the number-right score from the paper-and-pencil test (or, in some instances, a corresponding ability estimate) is based on a sum of unweighted item responses. Also, not reached items are treated very differently in CATs than they are in paper-and-pencil tests (see Way, Eignor, & Gawlick, 2001).

Psychometric characteristics of the CAT and paper-and-pencil tests also contribute to the lack of comparability of scores. With certain fixed-length CATs, the test length is set to yield the same level of overall reliability as the paper-and-pencil test in a representative group of examinees. Although this should ensure comparable overall standard errors of measurement (in that representative group), it in no way ensures that the conditional standard errors of measurement are equivalent. This will be true for comparisons made using observed scores (paper-and-pencil) and estimated true scores (CAT). This situation violates one of the assumptions of equating, the equity assumption, and, in particular, second-order equity (see Holland, Chapter 2; Kolen & Brennan, 2004). The equity criterion, in general, requires that it should be a matter of indifference as to which of two linked forms an examinee takes. This translates into very specific requirements about the level of precision to which scores on the two forms are measured. Second-order equity requires that examinees at a given ability level be measured with the same level of precision on the two test forms. In order for this to happen at a particular ability level, the conditional standard errors of measurement must be equivalent. The manner in which the fixed-length CAT was constructed will in no way ensure this is the case. Hence, with fixed-length CATs, second-order equity cannot be said to have been met.

With variable-length CATs, the length of the CAT is set to yield a specific level of precision. The CAT, however, is likely to provide greater precision than the paper-and-pencil test at any selected ability level.

Thus, the differing psychometric characteristics of the CAT and paper-and-pencil test lead to a lack of comparability of scores, such that the linked scores cannot be considered to be equated. This is why the term *calibration* has been used in this chapter to characterize the linking of scores on CATs and paper-and-pencil forms. After calibration, the sets of scores can be said to be equivalent in appearance only. Finally, on a very superficial level, there is no way that linked scores on a CAT and a paper-and-pencil form can lead to indifference on the part of an examinee as to which form she takes. Certain examinees will simply prefer to take the CAT, whereas others will prefer to take the paper-and-pencil form.

For a complete treatment of the issues leading to a lack of comparability between CAT and paper-and-pencil scores, the reader is referred to Wang and Kolen (2001). The issues discussed in the previous paragraphs are simply those that this author feels are the most important to emphasize.

8.4. Mode of Presentation Linking Designs

In the material that follows, taken in part from Eignor and Schaeffer (1995), the three most frequently used data collection designs for equating paper-and-pencil forms of an exam are discussed in the context of linking scores from computer-based and paper-and-pencil exams. These data collection designs are (a) the random groups design, (b) the single group counterbalanced design, and (c) the nonequivalent groups anchor test design. (See Kolen, Chapter 3, for detailed descriptions of these data collection designs.) Applications of these designs in a linking context will first be discussed, followed by a discussion of some modifications to these designs to deal with the peculiarities of computer-based administrations. These data collection designs will be discussed in the three linking contexts mentioned in Section 8.2: (a) equating linear CBT and paper-and-pencil scores, (b) calibrating CAT and paper-and-pencil scores, and (c) establishing a concordance relationship between linear CBT or CAT scores and paper-and-pencil scores where the computer-based test was not designed to be parallel to the paper-and-pencil test. Studies from the literature employing these data collection designs will be discussed where appropriate.

Table 8.1 provides a listing of the studies to be discussed, classified by data collection design and the type of linking employed. The studies are further broken down into those that attempted to demonstrate that the scores from the two modes of administration were equivalent and those that linked scores without checking their equivalence. Although no claim will be made that the studies listed represent the full set of studies that have been conducted, they are the studies that the author was able to locate, and they do provide an indication of the small amount that has been done to date in this area.

Before discussing these designs and related studies, it should be noted that the samples used in the linkings should be representative of the population to which the linking relationships will ultimately be applied. In the paper-and-pencil context, the samples should be representative of the population with respect to the distribution of the attribute being measured. With computers, other variables, such as level of computer familiarity, enter into the picture. In this case, the samples used in the linking need to be representative of the population with respect to both the attribute being measured and the level of computer familiarity or experience. This will be a particular issue when a CBT is to be introduced for the first time. Unless there are suitable practice materials and a viable tutorial, the examinees used in the linking study will likely not have the level of familiarity that examinees who take the test operationally will have. In this case, standards set on the CBT as a result of the linking study will likely demonstrate higher passing rates operationally than demonstrated in the linking samples. Determining appropriate linking samples in the computer-based testing context will necessitate a consideration of additional variables beyond those considered in the paper-and-pencil context.

Table 8.1. Summary of linking studies reviewed

		Data collection design		
		Random groups	Single group counterbalanced	Nonequivalent groups anchor-test
Equating	Equivalent scores established	Poggio, Glasnapp, Yang, & Poggio (2005)	Mazzeo, Druesne, Raffeld, Checketts, & Muhlstein (1991); Sykes & Ito (1997)	—
	Score equating performed	Schaeffer, Reese, Steffen, McKinley, & Mills (1993)	—	—
Calibration	Equivalent scores established	Eignor, Way, & Amoss (1994); Lunz & Bergstrom (1995)	Schaeffer, Steffen, Golub-Smith, Mills, & Durso (1995)	—
	Score calibration performed	Segall (1995); Segall & Carter (1995)	Eignor (1993); McBride et al. (1987)	Lawrence & Feigenbaum (1997)
Concordance	Equivalent scores established	Not possible	Not possible	Not possible
	Concordance tables produced	—	Jiang (1999)	—

8.5. Random Groups Design

8.5.1. General Discussion

One of the most frequently used data collection designs for studying whether scores from computer- and paper-based modes of administration are equivalent or for actually linking scores derived under these modes has been the random groups design. Such a design allows for straightforward statistical tests for differences in mean performance between groups across modes. Such hypothesis testing typically requires relatively large sample sizes. If the hypothesis test demonstrates no significant differences in mean performance across modes, then this might provide some initial indication that the two sets of scores are equivalent and linking can be viewed as unnecessary. To be certain that linking is unnecessary though, the test for mean differences in performance should be followed by a check of score distributions and score variances. Jaeger (1981) discussed how the Kolmogorov-Smirnov two-sample test of the equality of cumulative distribution functions (Smirnov, 1948) can be used to check for equivalent score distributions. Hanson (1996) discussed how log-linear models can be used to check on the equivalency of score frequency distributions. Finally, Segall (1995) discussed how an *F*-ratio test can be used to test for differences in score variances across the two modes of administration. If any of these statistical tests provide an indication that linking is necessary, the sample sizes needed to do the tests will prove useful because the random groups linking design requires relatively large sample sizes to keep linking errors at an acceptable level. (See Lord, 1950, for a discussion of these sorts of errors in the context of equating.) It should be mentioned that for all of the studies of this sort that have been reviewed in this chapter, only the test for differences in mean performance has been employed.

In addition to relatively large sample sizes, this data collection design requires a good deal of control over the examinees involved in the study. For instance, if the CBT is seen as an innovative form of assessment, examinees who have been randomly assigned to take the paper-and-pencil test might be disappointed and drop out of the linking study. Differential dropout is a major threat to this design because the two groups might no longer be comparable in ability. Hence, this design is better employed under conditions in which scores count, and the paper-and-pencil test provides a suitable avenue to attaining a valued outcome.

One distinct advantage of using a random groups design is that the same form can be given to both groups. In the case of equating a linear CBT to a paper-and-pencil test, each group would receive the same form via the different administration modes. In this situation, the linking does not need to take into account differences in difficulty due to different items. The linking must take into account only differences in difficulty brought about by administering the items via different modes.

In the case of calibrating scores on a CAT and a paper-and-pencil form, the above is not exactly true. If scores on the paper-and-pencil form used in the calibration have been reported on a scale separate from the raw score scale, as should be the case if there are multiple paper-and-pencil forms, then scores on the CAT have typically been reported on the same scale. This is done through use of a “reference form” that is part of the CAT system (see Eignor et al., 1993). The reference form has been previously given in a paper-and-pencil mode. In addition, items on the reference form will have been calibrated using the IRT model employed with the CAT and the item parameters placed on the IRT parameter scale for the bank. The initial score derived from the CAT will be an estimated ability score. Using this estimated ability and the item parameter estimates for the reference form, an estimated true score on the reference form can be derived. Estimated true scores on the reference form can then be linked with observed scores on the test given in the paper-and-pencil mode. One of the benefits of the random groups design is that it might be possible to administer the reference form from the CAT system in a paper-and-pencil format. So, as is the case with the linear CBT of the same content as the paper-and-pencil form in the equating context, the calibration of scores does not need to take into account differences in difficulty across test forms. However, it will need to take into account a much more expansive set of possible causes for differences in scores. Possible causes for these differences were discussed previously in this chapter.

Two final comments should be made about equating CBT and paper-and-pencil forms in the context of the random groups design. An advantage of this design in this context is that the same form can be given to both groups. This, however, does not need to be the case. Two different forms, say A and B, with A given as a CBT and B in paper-and-pencil mode, can be used instead. If A and B have previously been equated in paper-and-pencil format, then equated scores can be used for B in the subsequent linking across modes. In this situation, all that can differ across forms are the levels of difficulty caused by mode of administration. The above scenario does not seem to have been used in actual studies however. In the studies reviewed, Forms A and B have not previously been equated in paper-and-pencil format. Hence, the linking of A given via computer and B given via paper-and-pencil has to take into account differences in difficulty across forms due to both the use of different items on the forms and the use of the different modes of administration.

The other comment of relevance is related to how and when testing in the two modes can take place. In the context of equating two paper-and-pencil forms, randomization is usually brought about by packaging the test books in a spiraled order, and the two tests are administered simultaneously, usually in the same room. This is typically not the case when equating a CBT to a paper-and-pencil form; having computers in the same room where the paper-and-pencil test is taken could prove to be distracting. Hence, random assignment to conditions will need to be done in some other way than through spiraling, and the two groups will need to be separated for testing purposes. Also, in most situations, there will not be enough computers to test all of the examinees in the CBT group simultaneously. Instead, testing will need to be done over some time period. As long as additional learning does not take place during this time period for the CBT group, this arrangement would appear not to cause a problem. In fact, as will be seen in the discussion of the single group counterbalanced design in Section 8.6, this window of testing needed for the computer-based test can provide distinct benefits if testing is done properly.

8.5.2. Equating Studies Done with the Random Groups Design

Most times when researchers employ the random groups data collection design to study the comparability of scores from linear CBT and paper-and-pencil forms, a formal equating is not conducted. Instead, the same form is administered in both modes and statistical tests or informal checks of differences in mean scores are employed, and if the means are different, emphasis is placed on changing conditions under which the linear CBT is administered to ensure equivalent scores. For example, the administration of passage-based reading items on the computer might need to be altered so as to parallel to the extent possible the way such items are given in paper-and-pencil mode.

Mazzeo and Harvey (1988) provided a review of a large number of studies done prior to 1988 that employed the random groups design to study the equivalence of scores across linear CBT and paper-and-pencil administration of the same form. A wide variety of testing contexts are covered. The study by Schaeffer et al. (1993) to be discussed in Section 8.6.3 illustrates the practice of altering administration conditions to bring about equivalent scores. A more recent application of this data collection design with linear CBT and paper-and-pencil versions of tests can be found in Poggio et al. (2005). Finally, for a meta-analysis of some 30 studies employing the random groups design in studying scores on linear CBTs and paper-and-pencil tests, see Mead and Drasgow (1993).

In all of these studies, only differences in mean scores were looked at, and no attention was paid to possible differences in score distributions, which could imply that a formal equating study might still have needed to be conducted.

8.5.3. Calibration Studies Done with the Random Groups Design

A number of studies have been conducted that have employed the random groups design to look at comparability of scores between CATs and paper-

and-pencil tests. In certain of these studies, formal linking or calibration studies have not been conducted. Instead, IRT-based ability estimates have been derived from both modes of administration and their means directly compared. In other situations, calibration studies have been conducted to allow the derived scores from the two modes, which could not be directly compared, to be used (more or less) interchangeably. Following are some examples of both kinds of study. Eignor et al. (1994) employed the random groups design to look at whether sets of scores for the National Council of State Boards of Nursing Licensure Examinations (NCLEX), which were given in both computer-adaptive and paper-and-pencil modes, could be considered to be equivalent. Formal statistical hypotheses tests were conducted for differences in mean performance, using 1-PL IRT ability estimates, and for differences in passing rates, using log-linear models. In almost all cases, no significant differences in performance or passing rates were found. The study reported in Eignor et al. (1994) was somewhat unique in that the plan was to immediately replace the paper-and-pencil version with the computer version, and both exams were administered at the same time, with scores on both counting for licensure purposes. The focus of the study was on demonstrating that there would be no falloff in candidate performance with the switch in test modes.

Lunz and Bergstrom (1995) used a similar approach in one of a series of studies the authors conducted with the Board of Registry Certification Examinations for medical technologists. Examinees were randomly assigned to either CAT or paper-and-pencil conditions and equivalent mean performance across modes was taken as an indication that calibration of scores from the two modes of administration was not necessary. In this study, 1-PL model ability estimates were compared across the modes of administration.

It should be noted that in both the Eignor et al. (1994) and the Lunz and Bergstrom (1995) studies, no attempt was made to study the distributions of scores across the tests given in the two modes. Depending on the IRT model and calibration program employed, such a comparison might not be so straightforward. Also, even if means, variances, and distributions of scores could be established to be equivalent across modes, this would be in appearance only, as the scores would have different psychometric properties across the two modes.

For certain of the studies that employed the random groups design with CAT and paper-and-pencil versions of the same test, an actual calibration of the scores took place. Segall (1995) discussed the use of this design in linking scores on tests that are part of the Armed Services Vocational Aptitude Battery (ASVAB) and that involve computer-adaptive and paper-and-pencil counterparts of these tests. Segall and Carter (1995) discussed the planned use of the same design in calibrating scores on computer-adaptive and paper-and-pencil versions of certain tests that are part of the General Aptitude Test Battery (GATB).

In both contexts, an equipercentile procedure with smoothed frequency distributions was employed to bring about the calibration that created the conversion between scores from the two modes. What is noteworthy about both of these studies is that ability estimates from the CAT were linked directly to number-right scores on the paper-and-pencil form. It was not possible to transform scores from one of the administrations to allow a comparison of means or, for that matter, means, variances, and distributions.

8.6. Single Group Counterbalanced Test Design

8.6.1. General Discussion

With the single group counterbalanced design, all examinees take both the computer-based and paper-and-pencil versions of the test. Unlike with the random groups design, two separate tests must be used, that is, the computer and paper-and-pencil tests cannot be different versions of the same test form. In most early applications of this design, the tests were given sequentially in one testing session. A random half of the total group took the computerized test first and the remaining half took the paper-and-pencil version first. In this design, the first test taken might provide practice for the second test, thereby raising scores on the second test above what they would have been had the second test been given by itself. Fatigue might also lower performance on the test taken second. However, the relationship derived from the scores from the first administration of either version is what is of interest (i.e., the scores without practice or fatigue effects). This linking relationship could be estimated by ignoring the data from the second administrations and treating the first administrations as though they were obtained from a random groups design. However, the strength of the counterbalanced design is the potential to combine data from both administrations, thereby providing a much more precise estimate of the linking relationship than could be obtained from a random groups linking using the data from the tests administered first in each order. Another possible strength of this design is that it will likely be good for examinee motivation, given that the highest score is counted across the two testing opportunities, and everybody is provided the opportunity to take the test on computer. Both of these situations have the potential for helping with the possible dropout problem.

One key limitation of the counterbalanced design has to do with the conditions that must be met before the data can be combined and used. The procedure for combining counterbalanced design data makes some explicit assumptions about the nature of "order effects." An order effect in this context refers to the average change in scores, be it an increase (from practice) or a decrease (from fatigue), to be expected from the first administration to the second administration. The equations used for estimating equating parameters using all of the data assume that such order effects are in the same direction for both testing orders and are proportional to the standard deviations of the tests. The requirement that the average signed changes be equal in standard deviation units is usually difficult to meet in practice. It should be noted that the equations referred to are those in Angoff (1984); other counterbalanced linkings based on less restrictive assumptions have been discussed by Holland and Thayer (1990) and von Davier, Holland, and Thayer (2003), but have not been considered in the present context.

When the equations in Angoff (1984) are the focus and nonproportional order effects are present, then typically only data from the first administration of each test, treated as coming from a random groups design, can be used for linking purposes. Because the number of examinees in the counterbalanced design is usually small, in hope that the data from the orders can be combined, a linking based on only the first administration of each test will typically not be precise enough for operational use and additional data will need to be collected. See Kolen (Chapter 3, Section 3.5) for an additional discussion of these types of issue.

In more recently conducted linking studies using this design, the tests could not be given sequentially in one testing session because the availability of computers precluded the testing of all examinees in the computer mode at the same time. This situation has in many ways proven to be a blessing in disguise. If the timing between the two tests is such that there is no possibility of either practice or fatigue effects, then having two separate orderings of the versions of the test is no longer necessary. In most studies that have capitalized on this, the paper-and-pencil version of the test has been given first. The study done by Schaeffer et al (1995), discussed later in this section, is in this tradition.

Another possibility is to consider use of the full-blown counterbalanced design, but not worry about specific order effects. For instance, in the Eignor (1993) study, examinees were randomly assigned to a testing order. Paper-and-pencil testing was scheduled for the middle of the testing window. Examinees assigned to the computer-first condition could pick a specific day to test on the computer prior to the paper-and-pencil testing day, whereas examinees assigned to the paper-and-pencil-first condition could pick a specific day to test on the computer after the paper-and-pencil testing day. The testing window was established by considering how long a period between the first and second administrations could exist without being concerned that subsequent learning had taken place. Because testing in both modes was done on different days, practice and fatigue effects were viewed as being, for the most part, nonexistent, allowing data from the two modes to be combined. Although the above scenario would seem to be a viable way of collecting data with the counterbalanced design, two things happened in the Eignor (1993) study that caused difficulties: (a) test proctors at sites did not always use the rosters provided to randomly assign examinees to testing orders and (b) some test proctors chose to test certain examinees in both modes on the same day. Given all of this, a compromise was reached whereby linking was done separately in the two orders and then the separate linkings were averaged. In doing so, however, the advantage of being able to use the combined data to do a more precise linking was lost.

Finally, it is possible here to test for differences in the means across the test modes. However, unlike with the random groups data collection design, any statistical test applied in this context would need to take into account the repeated-measures nature of the data.

8.6.2. Equating Studies Done with the Single Group Counterbalanced Design

The studies that made use of the single group counterbalanced design in the equating context typically looked at whether sets of scores on the same form across modes could be considered to be equivalent rather than carrying out formal equating studies. In their review of earlier studies (i.e., prior to 1988) that compared linear CBT and paper-and-pencil versions of tests using the counterbalanced design, Mazzeo and Harvey (1988) found that order effects can be very different across orders, with such effects being considerably larger for the computer version when the paper-and-pencil test is administered first than vice versa. In these studies, however, the two ordered tests were always given sequentially in one testing session.

Mazzeo et al. (1991) looked at the comparability of computer linear and paper-and-pencil versions of the CLEP® General Examinations in Mathematics and English Composition. Because the number of available participants was small, the researchers chose to make use of a single group counterbalanced design. For a given sample size, greater precision in linking is gained from this design than from a random groups design, or from an anchor-test design, to be discussed later in this chapter. In the first round of data collection, Mazzeo et al. found the presence of order effects for both examinations. Modifications were made to the computer delivery system and then a second round of testing was undertaken with the two tests being given sequentially in one session. No order effects were found for English Composition so that the data could be pooled and average performance across the two modes could be compared. The differences in means were viewed as being nonsignificant, which implied that scores from computer administrations could be reported on the paper-and-pencil scale. In the case of Mathematics, order effects were still present and, hence, the data were not pooled for comparison of the means. Looking at means from only the first administration of each of the two tests, the differences were substantial. Rather than using this data to link the tests (sample sizes were extremely small), the authors suggested that further investigation and modification take place in an attempt to remove order effects.

Sykes and Ito (1997) employed a single group counterbalanced design to look at the equivalence of 1-PL model ability estimates across a linear CBT and a paper-and-pencil version of a licensure examination. When comparing the ability estimates, the authors found a significant order by mode of administration interaction effect such that there was a significant difference in ability estimates across modes when the paper-and-pencil form was administered first, but no significant difference in paper-and-pencil and computer-based ability estimates when the computer-based form was given first. In this study, the two tests were given sequentially in a single session. It is interesting to note that in the Mazzeo et al. (1991) study, the larger mean differences within order were found when the CBT was given first, but in both orders the test taken second had the higher mean. Hence, the Sykes and Ito results differ from the Mazzeo et al. results. The Mazzeo et al. results seem in part to be due to practice effects. Sykes and Ito hypothesized that their results had to do with examinee expectations of a positive experience taking a new CBT. When examinees received the CBT first, their expectations were immediately met and there was no later falloff in performance when taking the paper-and-pencil test. This was not true for the reverse ordering.

8.6.3. Calibration Studies Done with the Single Group Counterbalanced Design

There are three examples in the literature of calibration studies that made use of the single group counterbalanced design, or a variant of it. Two studies calibrated scores across CATs and paper-and-pencil forms. One study calibrated scores between a CAT and a linear CBT.

Schaeffer et al. (1995) used a variant of the single group counterbalanced design, where only a single ordering was used, to look at the comparability of scores from the CAT and linear CBT forms for the GRE®

General Verbal, Quantitative, and Analytical tests. In Schaeffer et al. (1993), the authors had established that scores were comparable across linear CBT and paper-and-pencil versions of these tests via a random groups design. (The ultimate goal here was to move the paper-and-pencil GRE General tests to CAT. The researchers chose to do this via a two-step process.) Each of the three CBTs, one for Verbal, one for Math, and one for Analytical, is given in two sections. Hence, to take all three CBTs, an examinee would end up taking six sections. Six scrambles, different orders of these sections, were created, and a Verbal, Quantitative, or Analytical CAT was given in the seventh or last position of each of these scrambles. (Two scrambles had the Verbal CAT, two had the Quantitative CAT, and two had the Analytical CAT.) An example of one scramble follows: V1 A2 Q1 V2 A1 Q2 VCAT. As can be seen, two sections of nonverbal material were given between V2 and VCAT, and this was true for all spirals. Although all sections were given in one session, practice effects were mitigated through the presence of two sections that contained different content prior to the CAT. This provides some justification for only using a single ordering where the linear CBT is always given first, followed by the CAT.

Schaeffer et al. (1995) used the results to create estimated true score to reported score conversion tables for the CATs and then compared them to the observed score to reported score conversion tables for the linear CBTs. For the Verbal and Quantitative CATs; these tables were viewed as being sufficiently comparable in nature that the linear CBT conversion tables could be used with the CATs. However, this did not prove to be the case for the Analytical CAT. Additional data were collected, via a “true” single group counterbalanced design, where the three GRE General CATs were given along with the Analytical linear CBT. The Analytical CAT was given first in one order, followed by the other two CATs, and then the Analytical linear CBT. In the other order, the Analytical linear CBT was given first, followed by the Verbal and Quantitative CATs, and then the Analytical CAT. It was found that, on average, the Analytical CAT scores were significantly higher than the Analytical linear CBT scores; hence, a calibration of scores on these tests was undertaken using an IRT true-score procedure. The linking results were then applied and the Analytical linear CBT reported score conversion was used with the Analytical CATs.

McBride et al. (1987) used the single group counterbalanced design to calibrate CAT and paper-and-pencil scores on selected tests from the Adaptive Differential Aptitude Test. Linear and equipercentile linking methods were employed. The equipercentile method was chosen in each case, because those results were superior to the linear linkings. In all linkings, the ability estimates on the CAT were linked directly to the number-right scores on the paper-and-pencil version. The authors did not discuss any analyses of scores in the two separate orders, and it appears that data were pooled across the orders. Hence, it must be assumed that order effects were not viewed as being a problem. Finally, it is not surprising that the equipercentile method was viewed as superior to the linear method with these linkings, given that number-right scores and ability estimates were used in the linking. For any test scored via IRT and then scored in a conventional fashion, the relationship between number-right and ability scores is nonlinear.

Finally, Eignor (1993) did a linking study between SAT® CAT prototypes in verbal and mathematics and their paper-and-pencil counterparts using the single group counterbalanced design. Many of the details of this study have been discussed in a previous section of this chapter. Noteworthy in this study is that the form used as the reference form in creating estimated true scores on the CAT was also used to generate the raw scores on the paper-and-pencil version. Eignor compared final raw to scale conversions for the CATs and the paper-and-pencil forms, where the CAT estimated true scores were linked to paper-and-pencil observed scores. Differences between the two conversion tables for particular raw scores were then scrutinized for verbal and for math. These conversions turned out to differ more than was expected, given the use of the same paper-and-pencil form to create scores. In retrospect, this is perhaps not surprising, given that the CAT reference form simply transforms ability estimates to a different metric. Differences between modes will still be evident after applying the transformation of the CAT ability estimates to the estimated true-score scale.

8.6.4. Concordance Studies Done with the Single Group Counterbalanced Design

A computerized version of the Test of English as a Foreign Language™ (TOEFL®) was planned in order to introduce new item types that took advantage of computer administration, add an essay to the Writing section, and change the structure of the Reading and Listening sections. All of these changes made the sections of the new test significantly different from the comparable sections of the old paper-and-pencil test. Consequently, a calibration of scores between the sections could not be considered. Hence, new scales were defined for each of the sections of the new test and also for the total score. After much discussion, it was decided that the Listening section and the Writing multiple-choice section of the new test would be CATs, whereas the Reading section would be what is referred to as a linear-on-the-fly test (LOFT; see Carey, 1999). Note that Kolen and Brennan (2004) referred to such tests as computer-based randomized tests.

A LOFT was chosen for Reading because it was felt that the level of item dependence among items within passages precluded the use of item-level CAT, and the details of testlet-based CAT had not been worked out at that time. Given that reporting scales were going to be discontinued for the paper-and-pencil test, there was interest in providing users with some idea as to where to set the cut-scores on the sections of the new test. Hence, it was decided that concordance relationships would be established to provide approximate cut-points on the computerized test sections that corresponded to the cut-points on the old paper-and-pencil test sections. The examinees in the study took the paper-and-pencil form at a TOEFL operational administration and then took the computerized form shortly afterward, in a nonoperational setting. Because order effects were expected to be minimal to nonexistent, only one order of the single groups counterbalanced design was used.

Aware that concordance relationships are particularly sensitive to the groups used to create them (Dorans & Walker, Chapter 10; Kolen & Brennan, 2004; Pommerich, Chapter 11; Pommerich & Dorans, 2004; Sawyer, Chapter 12), it was decided to estimate population score distributions on the computerized sections and use what could be assumed to be "representative of the population" distributions on the paper-and-pencil sections to create the concordance for each section. This represents the unique feature of this study and is documented in Jiang (1999). The paper-and-pencil population distribution was based on a national sample of 50,000 examinees that were representative of the complete population that had taken the paper-and-pencil test. The study sample of 7,057 examinees was a subset of the 50,000. From the paper-and-pencil population distribution along with the study-sample paper-and-pencil and computerized test distributions, an estimated population distribution was created for each computer-based test section. Then using the observed paper-and-pencil population distribution for each section along with the estimated population distribution for the corresponding CBT section and equipercentile linking, concordances were created and approximate cut-scores were provided for the new test sections. Actually, what is described above is a simplification of what was done in that the estimation of the computer population distributions was treated in a multivariate fashion and all section distributions were estimated simultaneously rather than one by one. However, it is useful to think of the estimation in the univariate context because of its similarity to frequency estimation observed-score linking (see Kolen & Brennan, 2004). It should be noted that all of this work was motivated by the belief that the use of the population distributions in the concordances would provide more appropriate computerized test section cut-points than if the concordances had been based on the distributions provided by the study sample of 7000+

examinees. One issue of concern though was whether the examinees in the study sample were sufficiently familiar with computer-based testing to adequately represent the group of examinees who would later be taking the computerized test in an operational setting. Even though a fairly extensive tutorial accompanied the new CBT, there were concerns about computer familiarity. In many of the other studies described in this chapter, computer familiarity does not seem to have been considered an issue.

8.7. Anchor Test: Nonequivalent Groups Design

8.7.1. General Discussion

An anchor-test design represents an alternative to use in lieu of collecting large examinee samples for the random groups design. In this context, an anchor-test design would involve two groups of examinees that are usually nonequivalent in ability. Under one possible scenario, one group would receive the CBT followed by the anchor test and the other group would receive the paper-and-pencil version followed by the anchor test. Under the other possible scenario, the anchor test would be given first in both groups, followed by the two tests for which scores will be linked. The anchor test could be a parallel form of the tests or it could be a shortened version of them. Additionally, the anchor test could be administered to both groups in either paper-and-pencil format or in computer-based format. Both groups would need to take exactly the same anchor test in exactly the same position and in exactly the same mode. It is this additional "wrinkle" that makes it difficult to conduct linkings when the test and anchor are to be given consecutively in one testing session. For one order, the test and the anchor would need to be given in different modes. Given this complication, this is not a design that would likely be considered in the linking of linear CBT and paper-and-pencil forms of an exam, where other designs work well. Given the necessity that the anchor test be parallel to both of the tests precludes the use of this design for test concordance purposes. Hence, this design would most likely be employed when calibrating scores on CAT and paper-and-pencil versions of an exam.

It should be noted that items could possibly be located to constitute an anchor test that operated in the same fashion regardless of the mode of administration. If this were possible, the anchor could be given via computer or via the paper-and-pencil mode. This is a design that has been employed in linking the scales of similar tests given in different languages, such as the English and Spanish versions of the SAT (Angoff & Cook, 1988). Note that the Spanish version was not a direct translation of the

English version; rather, the tests were constructed separately to test the same content. The anchor test in this case was made up of a separate set of items that were not part of either of the two tests. However, in this particular situation, nothing prevents the anchor items from being internal to the tests themselves. If this is the case, concerns about the influence that the tests have on the anchor when the anchor is given last or the influence that the anchor has on the tests if the anchor is given first likely become nonproblems. Further, although this scenario would seem to best hold when linking linear CBTs to their paper-and-pencil counterparts, if one is willing to seed the common items into the CATs, nothing would prevent this scenario from being used in the linking of CATs and paper-and-pencil versions of the test. Here everything hinges on establishing that the common items function in the same way when given in the computer and paper-and-pencil modes. This would need to be established prior to the linking study itself.

Like the single group counterbalanced design, the linking relationship of interest with this design is between scores on CAT and paper-and-pencil versions of the test that are uncontaminated by the effects of taking the anchor test. Given this, it makes some sense to give the anchor test after the tests for which scores are to be linked. Also, if the anchor test was given first, the possibility of nonequivalent practice effects exist. If it could be established that the groups were equivalent in ability, as might be the case if they were random groups from the same population, then when the anchor appeared last, it could be disregarded and the linking relationship estimated from the data from the two tests administered first. However, as with the single group counterbalanced design, the strength of the anchor-test design is that under certain conditions, the data from the tests to be linked and the anchor test can be used in combination to provide a more precise estimate of the linking relationship of interest than could be obtained using the random groups design with a comparable sample size (i.e., disregarding the anchor test).

The statistical theory behind the anchor-test design is based on some key assumptions regarding anchor-test performance. Specifically, the anchor test needs to be a comparable measure of the construct being assessed for both groups in the design. Scores on the anchor test must represent the same attribute being measured, apart from possible group differences in performance on that attribute. Such a condition implies that any order effects associated with the anchor test need to be the same for both groups. Thus, difficulties associated with the anchor test might not necessarily be circumvented by giving the anchor test after the two test versions for which scores are to be linked.

Possible order effects associated with giving the anchor test last or possible practice effects associated with giving the anchor test first become a nonproblem if the tests and the anchor are not given sequentially in one testing session. The time period between administration of the anchor and the tests to be linked would need to be such that no learning took place during the period. Also, if the anchor and the tests are given on separate occasions, the problem that the anchor will be given in a different mode than one of the tests in question also becomes a nonproblem. In fact, one study took advantage of just this sort of arrangement.

One possible concern about the anchor test and the tests to be linked being given on separate occasions is whether this causes any of the assumptions underlying the nonequivalent groups anchor test design to be violated. After all, when this design and the single group counterbalanced design are discussed in the literature, the treatments have either the test and the anchor, or the two ordered tests, administered contiguously in a single testing session. However, it is often the case in the context of equating forms of paper-and-pencil tests that the external anchor is administered at a different point in time than either of the tests to be equated. An example of this occurs with certain SAT II Subject Tests that are equated through external anchors consisting of SAT verbal and SAT math scores. Not only are the SAT scores from administrations at a different point in time than the SAT II adminstration, but the SAT scores themselves are from multiple different administrations. In sum, in the context of linking scores given in different modes, as long as the time period between the anchor administration and the test administration (or the two ordered test administrations) is such that no intervening learning of the test content can occur, noncontiguous administrations would appear not to cause problems with respect to underlying assumptions.

Finally, it should be noted that statistical tests of differences in performance across the two modes of administration are not possible with this design if the groups are nonrandom groups. If the groups are random in nature and the anchor is given last, the data from the anchor can be disregarded for hypothesis testing purposes. It is unclear what benefits could be derived from including the anchor items with the test items in doing statistical tests with random groups.

8.7.2. Calibration Studies Done with the Anchor Test Design

Lawrence and Feigenbaum (1997) used an anchor-test design to link a computerized-adaptive version of the SAT to the paper-and-pencil test. The SAT CAT and test linking described earlier in this chapter (Eignor, 1993) was never used for operational purposes. In the period between 1993 AND 1997, the CAT system was improved upon by, for instance, putting in appropriate item exposure controls. Also, the SAT was revised during this period of time. In addition, the College Board made a decision that an SAT CAT would be used operationally with students who were seeking placement into talent search programs such as the Johns Hopkins Center for Talented Youth. Hence, it was felt that a linking study needed to be done with scores from the new SAT CAT system prior to operational implementation. The linking or calibration study was done using data from regular SAT examinees, although the test was later to be targeted for talented youth.

To conduct the study, the researchers identified a group of examinees who had taken a paper-and-pencil SAT form at an operational SAT test administration. A subset of these examinees was invited to take a CAT version of the test 1 month later; those who subsequently took the CAT formed one of the nonrandom groups. The other nonrandom group was made up of those examinees from the original group who took another operational paper-and-pencil SAT 1 month later. Scores for these two groups were calibrated making use of an anchor-test design where the score from the original operational SAT administration was used as the anchor score. Distributions were smoothed via log-linear procedures and then the chained equipercentile method (see Kolen & Brennan, 2004; Kolen, Chapter 3) was used to link the scores from the CAT and the paper-and-pencil test. New conversion tables were created for the CAT and compared to the paper-and-pencil conversion tables that existed for the paper-and-pencil form taken in the second administration. (As described earlier, with the CAT, the estimated abilities were transformed into estimated true scores on a reference form that already had a raw to scale conversion table.) The magnitude of the differences in the conversion tables for both verbal and math ranged from 0 to 20 scaled-score points. It is interesting to note that the magnitude of differences from the Eignor (1993) study, done using a different CAT system and data collection design, were also between 0 and 20 scaled-score points, and the nonzero differences in the conversion tables were in the same relative spots on the raw score scales in both studies. However, although the differences in the two studies were in the same direction for verbal, they were in opposite directions for math. Lawrence and Feigenbaum (1997) had a number of concerns about their study, including possible differential motivation levels between the examinees who took the CAT and those who took the subsequent SAT, for which scores were reported as usual. However, it should be noted that the examinees who took the CAT were given the option of keeping their scores or having them canceled. This should have helped to eliminate, to some extent, possible motivational differences.

8.8. Summary

A number of issues surrounding the linking of scores on computer-based and paper-and-pencil tests were discussed in this chapter. The linking relationship between the tests could be characterized as being an equating, a calibration, or a concordance, depending on the nature of the CBT. The linking for which the most issues surface involves scores on CAT and paper-and-pencil forms of a test. Although users would like to be able to use the scores from these two sorts of test interchangeably, as would be the case if the linking of these scores could be viewed as an equating, such a linking can be considered to be, at best, a calibration, primarily because it cannot be shown to satisfy the equity requirement of equating. This, however, has not stopped score users from treating calibration results as if they were equating results. In fact, because the calibration process typically yields scores that are reported on the same scale, it is only logical that users will treat the calibrated scores as if they were equated scores. At present, it is not exactly clear what the consequences are of treating a calibration in this context as if it were an equating. It might very well prove to be the case that such scores should be related via a concordance table. Separate scales would exist for the forms and there would be less inclination to use the results as if they came from an equating. However, unless specifically cautioned, users will often use the scores related via a concordance table as if they were equated scores, even though the scales themselves will likely be different. Pommerich (Chapter 11) expressed similar concerns with concordance tables.

The population invariance requirement for the equating of scores was not specifically discussed in this chapter. However, a number of reviewed studies took a look at the effects of linking transformations on subgroups. These include the studies done by Poggio et al. (2005), Schaeffer et al. (1993, 1995), Eignor et al. (1994), Segall (1995), and Lawrence and Feigenbaum (1997). In all of these studies the (sub)population invariance property was not specifically investigated because separate subpopulation linkings were not undertaken. This is partly because the invariance checking procedures (see Dorans & Holland, 2000) had not been developed at the time that most of the studies were done. It would be particularly interesting to see whether linkings between CATs and paper-and-pencil tests, which have been shown not to satisfy the equity requirement, also do not satisfy the population invariance property. This would add additional strength to the assertion made in this chapter that the linking between scores on a CAT and a paper-and-pencil test does not qualify as an equating.

A significant portion of the chapter dealt with data collection designs necessary to link scores on CBTs and paper-and-pencil tests. This was done in hope that the chapter might provide some guidance to other practitioners faced with similar linking situations. It was seen that the random groups design provides a mechanism for linking scores on these tests that is basically free from the influence of practice or fatigue effects, or order effects in general. It was also shown, however, that nontraditional applications of the other designs, whereby tests to be linked are not given consecutively in one testing session, also provide viable options for linking scores.

One final caveat is in order. Much of the material in this chapter was based on the author's personal experiences in linking computer-based and paper-and-pencil forms of tests and on the experiences documented in 11 articles located in the literature. This is clearly too small a set of articles on which to draw general conclusions of any sort, and it might prove to be the case when further studies are conducted that certain of the conclusions in this chapter might need to be altered. In fact, this has already happened, given that in earlier work this author and other authors considered the linking of scores on CATs and paper-and-pencil tests to qualify as equatings, whereas more recent work, such as the work of Wang and Kolen (2001), has shown this not to be the case. Finally, it should be pointed out that the number of studies in the literature addressing linking of this sort, and the related problems, will never be voluminous in nature because testing programs most often do these sorts of study only once, or perhaps a small number of times, as the programs are transitioned from paper-and-pencil to computer-based testing.

Acknowledgments. The author thanks John Mazzeo and Lin Wang for their reviews of a draft version of this chapter.