11

Scaling, Equating, and Linking

Tim Moses
Buros Center for Testing

Moses, T. (2025). Scaling, equating, and linking. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 735–820). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0011

> Few wish to assess others,
> Even fewer wish to be assessed,
> But everyone wants to see the scores
> (Paul Holland, quoted in Dorans, 2018)

There are many ways that a testing program can assign and report scores that reflect test performance. Score reporting can be based on performance standards set by content experts; on relative standing with respect to one or more test-taker groups; on performance, strengths, and weaknesses across a set of tests contained in a test battery; or on performance for a particular age group or grade level in a scale constructed to track growth in age- or grade-appropriate tests of increasing difficulty. It would be difficult to overstate the importance of reported scores for a large-scale testing program. As a testing program's most visible and widely used products (Dorans, 2002; Kolen, 2006; Zenisky et al., this volume), it is essential to understand and document how reported scores are produced, maintained, and, at times, related to the scores reported by other testing programs. This chapter focuses on procedures for producing, maintaining, and linking testing programs' reported scores.

This chapter is an updated version of prior writings on scaling, equating, and linking. The attempt is to build on the seminal and foundational work of the chapters in previous volumes of *Educational Measurement* (Angoff, 1971; Flanagan, 1951; Holland & Dorans, 2006; Kolen, 2006; Petersen et al., 1989). Extensive reviews of these chapters are, regrettably, not possible here, but are strongly encouraged. The major goal of the current chapter is to provide a "point-in-time" description of how issues and problems in the current testing field engage with, challenge, and build on prior chapters' discussions and frameworks for scaling, equating, and linking a testing program's reported scores.

- *Scaling* focuses on the theory and practice of establishing a testing program's reporting score scales.
- *Equating* focuses on the definition, requirements, history, and methodological practices of equating for maintaining a testing program's reporting score scales.
- *Linking* types are summarized with historical and recent examples for how a testing program might relate its reported scores to other scales.

In addition to the scaling, linking, and equating chapters in previous *Educational Measurement* volumes, the current chapter reflects two major influences. First, the testing program from which reported scores are produced represents a system of test production, administration, scoring, using, and interpreting test results, in multiple testing sites, and repeating in cycles over multiple points in time (Dorans, 2011; Holland, 1994, 2008). Understanding reported scores requires an understanding of the testing program's system, especially test specifications and development, administration conditions, and scoring procedures. The second influence pertains to the interpretations of reported scores. Testing standards emphasize that reported scores should be produced and maintained in ways that encourage appropriate interpretations and discourage misinterpretations (American Educational Research Association [AERA] et al., 2014; Kolen, 2006; Petersen et al., 1989). The discussions in this chapter are intended to specify and elaborate on what this means and on ways to most effectively reflect "the overarching consideration . . . that users be given appropriate guidance about score interpretation and use" (Brennan, 2007, p. 175).

SCALING

The purpose of scaling is to establish the reporting scale(s) for the individual measures of a new or redesigned testing program. Scales are produced in the context of several activities of a testing program, such as the development of a test or measure from an established set of specifications for the content, construct, and individual items. Scaling is used in several testing contexts, including:

- admissions testing, where testing companies develop tests with relatively long scales for use in admissions decisions;
- K–12 , where testing companies develop content-based tests associated with specific curricular standards adopted by policy makers to classify students at one or very few cut points for school and educator accountability, high school exit requirements, and instructional decisions;
- certification and licensure programs, where test content is determined with extensive input from practitioners and test takers are classified at cut points that indicate sufficient knowledge and skills to qualify for professional practice; and
- large-scale survey assessments developed by policy makers, educators, measurement and content experts, and other stakeholders for use in the estimation of trends in scale score distributions and classifications of populations and subpopulations over time.

Scaling activities also involve the administration of the developed test to test takers from a defined testing population under specific administration conditions, the scoring of test takers' test items, and the conversion of test takers' item scores into overall test scores or ability estimates. The data collection design for a test-taker sample from population P who take test form Y is shown in the first row of Table 11.1 (the terms and other designs in Table 11.1 are described throughout this chapter). The task in scaling is to develop a transformation that assigns numbers or ordered indicators to the test performance data.

The scaling process reflects the tests and their intended measurement, interpretations, and perspectives reviewed in the "Scaling Perspectives" section. Consider Y as a test form and a set of performance values test takers might receive, such that specific performance values are denoted $y = 0, 1, 2, \ldots$. The ys can reflect sums of correct item scores, weighted summed scores, summed predicted probabilities of item-level models, or other indicators as described in the *Basic Unit of Scaling* section. Scale score transformations of

Table 11.1 Summary of Data Collection Designs for Scaling, Equating, and Linking

Description	**Design Table**					
Scaling/norming for a single test	Population	Sample	Y			
	P	1	√			
Single group	Population	Sample	X	Y		
	P	1	√	√		
Randomly equivalent groups	Population	Sample	X	Y		
	P	1	√			
	P	2		√		
Counterbalanced	Population	Sample	X_1	Y_1	X_2	Y_2
	P	1	√			√
	P	2		√	√	
NEAT	Population	Sample	X	A	Y	
	P	1	√	√		
	Q	2		√	√	
Common-item equating to a calibrated pool	Population	Sample	X_A	X_{New}		
	P	1	√	√		
Section pre-equating	Population	Sample	X_1	X_2	Y_1	Y_2
	P	1	√		√	√
	P	2		√	√	√
Item pre-equating	Population	Sample	X	*New*		
	P	1	√	√		

Note. NEAT = nonequivalent groups with anchor test. Depending on the context, X and Y can refer to alternate forms of a single test (equating) or to distinct tests (linking). For all but the NEAT design, P is assumed to be the target population for the test (T).

test performance are developed to produce reported scale scores that reflect increasing levels of achievement or proficiency, $sc(Y)$, that encourage specific interpretations. Two types of scales might be developed from the inputs, *primary* and *auxiliary*. Primary scales refer to reporting scales and test performance indicators that underlie the psychometric operations of a test (Kolen, 2006; Petersen et al., 1989), including the equating of scores from alternate test forms (see the *Equating* section) to the primary scale. Methods of establishing primary scale scores are described in the section *Scaling Methods for Primary Scale Scores*. Auxiliary scales are developed to facilitate interpretations and convey meaning in primary scales, and are described in the section *Auxiliary Scales*.

Scaling Perspectives

Early antecedents for educational and psychological testing and scaling include Horace Mann's efforts in 1845 to standardize oral exams of test-taker placement to control for the influence of different topics, examiners, and other efforts in psychophysics to apply scientific methods to study relationships between psychological sensations and physical stimuli (Briggs, 2022; Mislevy, 2018). Efforts to establish scales for nonphysical, psychological measures reflected these antecedents, as well as aspirations that measures and scales of nonphysical attributes might exhibit the measurement and scaling properties of physical attributes. The Binet–Simon Intelligence Test, released in France in 1905 by Alfred Binet and Theodore Simon, was an attempt to identify children in need of special services in ways that would be objective and free of teacher judgment and variations in the terminology, evidence, and reasoning used at that time (Binet & Simon, 1916; Cronbach, 1949). In the Binet–Simon test, test takers between ages 3 and 13 completed a series of 10 to 30 age-specific tests and were given a scale score designating an age and "intellectual level" corresponding to the test they were able to complete (Becker, 2003; Cronbach, 1949):

> The fundamental idea of this method is the establishment of what we shall call a measuring scale of intelligence. This scale is composed of a series of tests of increasing difficulty, starting from the lowest intellectual level that can be observed, and ending with that of average normal intelligence. Each group in the series corresponds to a different mental level. (Binet & Simon, 1916, p. 40)

In another early example of scaling and measurement, Thorndike (1910) proposed a scale for handwriting quality by obtaining rankings of subsamples of handwriting samples of children in the fifth to eighth grades from competent judges and averaging these rankings to produce a scale for the entire sample. Thorndike (1910) described his scale of handwriting quality as one that ranged from "better than which no pupil is expected to produce, down to a quality so bad as to be intolerable, and probably almost never found, in school practice in the grammar grade" (p. 89). He also considered his scale of handwriting quality a way to assess educational outcomes similar to physical and scientific measures in other disciplines:

> In general, the experience of constructing this scale gives great encouragement to the hope that for many educational facts, units and scales may be invented that shall enable us to think quantitatively in somewhat the same way that we can about facts of physics, chemistry or economics. (Thorndike, 1910, p. 150)

Scaling, measurement processes, and models were developed beyond the work of Binet and Simon (1916) and Thorndike (1910). For the Binet–Simon tests (i.e., items or questions) that were designed to be taken by children of specific ages, Thurstone (1925) developed a method to place the statistics of all Binet–Simon tests onto a single, continuous, normally distributed ability scale. This scale allowed Thurstone to evaluate and critique the a priori age groupings of the tests: "The questions are unduly bunched at certain ranges and rather scarce at other ranges" (Thurstone, 1925, p. 448). Thurstone (1926, 1927) also provided a rationale and more fully developed theory of Thorndike's (1910) scale in his law of comparative judgment (Coombs et al., 1970). This allowed for scales to be produced from any process of pairwise comparisons through modeling averages of pairwise differences and assuming they followed normal distributions. Guttman (1944) proposed a scaling approach that predicts an individual's agreement/disagreement with an ordered set of unidimensional attitude statements in a deterministic model asserting that individuals who agree/disagree with a stronger/milder statement of attitude will also agree/disagree with all milder/stronger statements. Rasch (1960) proposed a scaling procedure based on a statistical logistic item response model of the probability of an individual's response to an item based on one parameter for each item's difficulty and one parameter for an individual's ability (see the *IRT Ability Estimates* section; Wright, 1977).

Despite the methodological differences in previously described scaling approaches, comparative summaries have pointed out several similarities. The scaling approaches of Thorndike (1910) and Thurstone (1926, 1927) reflect an assumed "monotonicity property" in the nonphysical variables being scaled, implying consistency in judgments and subjective differences between stimuli (Coombs et al., 1970, p. 42). Thurstone's, Guttman's (1944), and Rasch's (1960) approaches assume invariant comparisons in scales intentionally constructed to be unidimensional (Andrich, 1988). Engelhard (1984) compared the item calibration approaches of Thurstone (1925) and Rasch (1960) and another approach from Thorndike (1919), showing that these approaches produce similar empirical results and that they all attempt to eliminate the effects of samples and groups and thus achieve invariance. These scaling approaches can be described as giving increasingly greater emphasis to a particular model over data (Engelhard, 1984, p. 35). An implication of these approaches' emphases on models is that uses and interpretations of the scales occur "only after a scale is developed that adequately fits the model" (Kolen & Brennan, 2014, p. 373).

Scaling has been addressed in fundamental definitions of what constitutes a measure and its scale. Early work was based on a classical perspective of discovering or estimating quantities and numerically representing them (Campbell, 1928; Hölder, 1901; Michell, 2008). From this perspective, attributes are established as quantitative and fundamental measures to the extent that they exhibit numerical properties that have a physical analogue, like length and concatenation (the adding or laying of objects end to end, Coombs et al., 1970; Hölder, 1901; Mislevy, 2018). An implication of these arguments is that measurement can be established only in the physical sciences. Psychophysical variables could not be established as measures because psychophysical variables would not exhibit concatenation operations or invariant relationships with established physical measures (Campbell, 1928; Ferguson et al., 1940).

Stevens (1946) provided a general typology of scales achieved through rules of assignment of numerals to a wide range of objects or events. By focusing on assignment rules rather than on requisite conditions for measurement (e.g., Campbell, 1928), Stevens's (1946) scaling approach is more encompassing, alternatively described as paraphrasing (Stevens, 1946), circumventing (Mislevy, 2018), shifting (Briggs, 2022), or deflecting (Michell, 2008) the emphasis of earlier arguments on classical measurement and quantitative measures. For Stevens, scales are achieved in an operationalist (vs. classical) process of following (vs. discovering) rules for assigning numerals to objects according to their assumed level of measurement (vs. attributes of objects that must first be established as measurable; Briggs, 2022). The objects and measurements for which scales can be established include nominal (categorical labels), ordinal (ordered labels), interval (ordered labels that have equal intervals), and ratio (interval scales with a natural zero). In Stevens's arguments, the meaning and interpretation of scales depends on using admissible transformations for types of measurement (i.e., a monotonically increasing transformation for ordinal, linear transformations for interval) and on using permissible statistics for specific types of scales (e.g., medians for ordinal scales, means for interval scales). Stevens's scaling theory was further developed in subsequent work (Coombs et al., 1970; Krantz et al., 1971; Suppes & Zinnes, 1963).

Other approaches re-emphasized the classical focus on quantitative (interval or ratio) measurement through conditions other than concatenation (Coombs et al., 1970), including conjoint measurement (Luce & Tukey, 1964). One condition of conjoint measurement is double cancelation, which refers to a variable that consistently, additively, and noninteractively increases or decreases with two other variables (see the *Ability Estimates From Rasch and Other 1PL IRT Models* section).

Scaling approaches used in practice and described in measurement theory have been nearly separate pursuits in educational and psychological testing: "The axiomatic analysis of measurement models does not always provide feasible methods for constructing scales" (Coombs et al., 1970, p. 31). Binet and Simon (1916) described inconsistencies in the actual and aspirational measurement properties of their Binet–Simon test, simultaneously arguing for treating the scale of their test as equal to a quantitative measure even as they acknowledged that it reflected an ordinal scale of ordered, discrete classes:

> This scale properly speaking does not permit the measure of intelligence, because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured, but are on the contrary, a classification, a hierarchy among diverse intelligences; and for the necessities of practice this classification is equivalent to a measure. We shall therefore be able to know, after studying two individuals, if one rises above the other and to how many degrees, if one rises above the average level of other individuals considered as normal, or if he remains below. (pp. 40–41)

Another argument for treating ordinal measures as though they reflected interval scales in scaling practice appealed to pragmatics and usefulness over formal measurement properties:

> Although, formally speaking, interval measurement can always be obtained by specification, such specification is theoretically meaningful only if it is implied

> by the theory and model relevant to the measurement procedure. At various times in this book however, we shall treat a measurement as having interval scale properties, although it is clear that the measurement procedure and the theory underlying it yield only a nominal or, at best, an ordinal scale . . . from a pragmatic point of view, the only meaningful evaluation of this procedure is one based on an evaluation of the usefulness of the resulting scale. (Lord & Novick, 1968, pp. 21–22)

Some objections have been expressed about Stevens's (1946) arguments concerning levels of measurement and permissible statistics, such as Lord's (1953) argument that statistical operations seemingly reserved for interval scales could be appropriate even for nominal scales like numbers on football players' jerseys. Connections of measurement theory and scaling practice have been considered in terms of the statistical and probabilistic Rasch models and conjoint measurement (Embretson & Reise, 2000; Karabatsos, 2001; Perline et al., 1979). However, noted limitations in the fit of the Rasch model have raised questions about the inconsistencies of pursuits to establish scales that reflect measurement theory versus those that fit actual test data (Briggs et al., this volume; Embretson, 2006; Embretson & Reise, 2000; Thissen & Orlando, 2001; Wright, 1994).

The perspectives taken in describing the practice of establishing reporting scales for educational tests in the first four volumes of *Educational Measurement* make specific references to, and departures from, previous discussions of scaling and measurement (Angoff, 1971; Flanagan, 1951; Kolen, 2006; Petersen et al., 1989). Most of the *Educational Measurement* chapters present a basic definition of scaling similar to Stevens's (1946), one of assigning numerals or numbers to a test taker's test performance. Interpretation and meaning are less about requirements for measurement, statistics, or scaling transformations and more about the scale properties that might be desirable for other criteria and interpretations (described in the *Scaling Methods for Primary Scale Scores* and *Auxiliary Scales* sections). All of the *Educational Measurement* volumes discussed scale score interpretations in terms of scale score distributions on relevant testing populations (norms). Discussions of interval scales and other measurement issues either appear in other nonscaling chapters (e.g., Lorge, 1951) or are described in terms of their limited usefulness for scaling and psychological measures due to the absence of operational definitions that are agreed on by experts (Angoff, 1971) and the lack of a complete theory and definition for the psychological and educational constructs in need of scales (Kolen, 2006). Finally, perspectives on scaling for educational tests emphasize a starting point that differs from the one implied in measurement definitions and models. Petersen et al. (1989), Kolen (2006), and Brennan (personal communication, September 17, 2020) have distinguished educational measurement from other types of measurement because the starting point in educational measurement, that is, test content and specifications, comes from an external entity (e.g., one of the testing contexts summarized in the beginning of *Scaling*). They argue, consistent with Lindquist (1953), that

> the objective is handed down . . . by those agents of society who are responsible for decisions concerning educational objectives, and what the test constructor must do is to attempt to incorporate that definition as clearly and as exactly as possible. (p. 35)

For educational measurement, content, test development, and measurement models influence the development of a test's reporting scale in choices for the basic unit used in scaling (see the *Basic Unit of Scaling* section), transformations of test performance measures to a reporting scale with desired properties (see the *Scaling Methods for Primary Scale Scores* section), and maintenance of the scale in subsequently developed forms through equating (see the *Equating* section).

Basic Unit of Scaling

The *Scaling Perspectives* section described fairly different views on scaling. These differences are apparent in the options for defining the basic unit of test performance used to establish a testing program's score scale. This section covers choices and implications for representing test performance, including different types of observed test scores that summarize observed test performance, and the latent abilities estimated using some item response theory (IRT) models. For other summaries of these options, including descriptions of approaches for scoring different types of items, see Kolen (2006) and Dorans (2018).

Observed Test Scores

Test performance indicators can be established as different types of observed scores. For test form Y that contains $i = 1$ to I items that are not distinguished with respect to content area or item format, the most common observed score is the sum of the $i = 1$ to I item scores, V_i,

$$Y = \sum_i w_i V_i \tag{1}$$

where $w_i = 1$ for all I items. With this option, a test taker's score reflects equally weighted contributions from each item in the test. Other options involve the summing of differentially weighted item scores, where the item score weights might be chosen to maximize some measure of test reliability or validity (Gulliksen, 1950; Lord & Novick, 1968). Although these reliability and validity measures for maximization are defined by Gulliksen (1950) and Lord and Novick (1968), many issues concerning reliability, true score variance, and error variance for scaling procedures are, at best, imperfectly understood. Other texts should be consulted for comprehensive discussions of reliability (Brennan, 2001; Haertel, 2006; Lee & Harris, this volume). Different options suggested in Equation 1 can be used to summarize test takers' test performance on a specific set of test items.

Test form Y may also be a "composite" or "mixed format" test containing items with different content areas and/or item formats. When Y includes I multiple-choice (MC) items and *J* constructed-response (CR) items, an approach to weighting the item type scores and obtaining observed composite scores is

$$Y = w_{MC}\sum_i V_{i,MC} + w_{CR}\sum_j V_{j,CR}. \tag{2}$$

One option is to nominally weight the scores from the MC and CR items so that the weighted scores reflect an intended number and percentage of composite score points. The Advanced Placement exams provide several simple and complex examples of MC item score weights and CR item score weights being set such that the weighted and summed scores contribute desired numbers of points to the composite score (College Board,n.d.; Moses et al., 2006). Another option is to determine effective weights such that the weighted scores reflect a desired proportion of the composite's observed score variance or true score variance. A third option is to select weights for the scores summed from different item formats such that some measure of composite score reliability is maximized.

The options for summing item scores produce measures of test performance with different implications. Sums of item scores that are equally weighted or weighted to reflect intended numbers of composite score points are the simplest methods of reporting test performance, associated with scoring rules that can be fairly easily communicated to test takers (i.e., "To do as well as possible, answer every item"). Simple summed scores can directly reflect expert judgments about the test as indicated by the numbers and weights of different items described in test specifications. Other options weight items based on statistics such as variances, maximized reliabilities, or validities and have interpretations based on those criteria. In addition, the weights reflect population and sample characteristics of those corresponding statistics. The scoring rules associated with these options for weighted observed test scores are more complex and may be more difficult to communicate to test takers depending on when these population-dependent weights are derived (i.e., test takers may approach the test in a way that is suboptimal or inconsistent with item weights that may be derived before or after the test is administered). Most of these observed score options should be understood to reflect an unspeeded test where test takers have sufficient time to answer every item. Inadequate testing time can elicit rushed or random responding that could result in unintended measurement properties for those speeded items and the overall test. Equally or nominally weighted item scores may be suboptimal with respect to test score reliability, especially if these weights result in higher contributions from the scores of less reliable item types (i.e., higher contributions of CR scores relative to MC scores).

Observed test scores are direct indications of actual test performance and indirect reflections of unobservable latent abilities. As such, potential uses of observed test scores as estimates of latent abilities require supplemental measures of the reliability or generalizability of the scores to other possible item samples, test administrations, and admissible measurement conditions. When interpreted in terms of measurement and scale properties, test scores are often described as ordinal scales for some latent ability they are assumed to measure (Embretson & Reise, 2000). Alternatively, observed test scores could be considered interval scales from a pragmatic perspective

(Lord & Novick, 1968), or from a perspective where scores from a particular test form are regarded as indices of test performance rather than estimated abilities, such that a test form's scores exhibit equal increases with additional correct responses to test items (Mislevy, 2018, p. 316).

IRT Ability Estimates

Scales might be established on estimated latent abilities rather than on the test performance observed for a set of test items. In theory, the item response data from a test form might be modeled in ways that produce estimates of abilities that are independent of the test form. In practice, this involves fitting IRT models to data from a test and its fixed, specific set of items, such that the item characteristics require scaling procedures to account for their sample effects (described by Kolen & Brennan, 2014). The result is that test performance measures can be produced that may be regarded as estimates of latent ability, but that are essentially computationally complex summaries of observed item performance, in need of their own reliability estimates (Lee & Harris, this volume).

Assume test form *Y* contains dichotomously scored items with correct responses scored 1 and incorrect responses scored 0. Assume further that test takers' responses to the *I* items are conditionally independent given a latent and unidimensional ability, θ, and that the probability of a correct response follows a logistic model with up to three parameters, such that the probabilities monotonically increase with θ. The item response probabilities from the one-, two-, or three-parameter logistic models (1PL, 2PL, 3PL) can be expressed as

$$Pr\left(V_i = 1|\theta, a_i, b_i, c_i\right) = c_i + \left(1 - c_i\right)\frac{1}{1 + exp\left[-Da_i\left(\theta - b_i\right)\right]} \tag{3}$$

where a_i, b_i, and c_i are discrimination, difficulty, and guessing parameters for item *i* and *D* is a scaling constant that is sometimes 1 and other times set to 1.702 so that Equation 3's logistic function approximates a normal ogive (Haley, 1952). This section summarizes some approaches for estimating test takers' abilities, θ, based on test takers' scored responses to the items on test form Y, v_1, v_2, ... v_I and on parameter estimates for all *I* items on the test form that are treated as population values (not estimated). IRT ability estimates reflect several choices for fitting IRT models to test data, including calibration decisions, estimation software, and approaches for polytomously scored items and mixed format tests (for additional discussions of IRT models, see Cai et al., this volume).

Three commonly used IRT ability estimates are the maximum likelihood estimate (MLE), the Bayesian expected a posteriori (EAP), and the test characteristic curve (TCC) estimate (for others, see Thissen & Orlando, 2001).

The MLE is obtained by solving for θ using an iterative procedure to maximize,

$$\begin{aligned} &L(V_1 = v_1, V_2 = v_2, \ldots V_I = v_I \mid \theta) \\ &= \prod_i Pr(V_i = 1|\theta, a_i, b_i, c_i)^{v_i}[1 - Pr(V_i = 1|\theta, a_i, b_i, c_i)]^{(1-v_i)}. \end{aligned} \tag{4}$$

The MLE can be described as the ability estimate with maximum information (i.e., precision) and inverse sampling variance. The MLE corresponds to an optimally weighted observed score, $\sum_i w_i V_i$, for a particular IRT model (Lord, 1980; Yen & Fitzpatrick, 2006).

The Bayesian EAP estimate,

$$EAP = \frac{\int_\theta \theta L(V_1 = v_1, V_2 = v_2, \ldots V_I = v_I \mid \theta) Pr(\theta) d\theta}{\int_\theta L(V_1 = v_1, V_2 = v_2, \ldots V_I = v_I \mid \theta) Pr(\theta) d\theta}, \tag{5}$$

is the mean of a distribution of θ obtained as the product of the likelihood and an assumed prior distribution for test-taker ability, $Pr(\theta)$. In practice $Pr(\theta)$ is usually represented as a discrete approximation of the standard normal distribution, though it can also be obscure.

The TCC estimate is based on relating the IRT-expected summed scores given θ, $\tau_Y(\theta)$, to the observed summed score (Equation 1). Ignoring measurement error in the scores of test form Y and making other adjustments to account for true scores that are undefined in the IRT model, Equation 1 and a summed version of Equation 3 are set equal to each other and θ is estimated to preserve this equality:

$$Y = \sum_i w_i V_i = \sum_i w_i Pr(V_i = 1 \mid \theta, a_i, b_i, c_i) = \tau_Y(\theta) \tag{6}$$

Although the solution to θ in Equation 6 ensures that the IRT-expected summed score matches the observed score, this application of true score relationships to observed scores lacks justification (Kolen & Brennan, 2014, p. 201).

ABILITY ESTIMATES FROM RASCH AND OTHER 1PL IRT MODELS For the Rasch (1960) (all $a_i\text{s} = 1$, all $c_i\text{s} = 0$, and $D = 1$) and other 1PL models (all $a_i s = constant$, all $c_i\text{s} = 0$, and D =1.702), ability estimates have a one-to-one relationship with the test score summed from equally weighted items, all $w_i\text{s} = Da_i$, and the TCC and MLE estimates are equal (Thissen & Orlando, 2001). When a uniformly distributed prior is used, $Pr(\theta)$, EAP ability estimates are also equal to those of MLE and TCC, except that the highest and lowest summed scores have estimates from EAP but are undefined with MLE and TCC. These results reflect a property of Rasch and 1PL models that the summed test score with equal $w_i\text{s}$ is a sufficient statistic that contains all information needed for estimating θ (Lord, 1980). Response probabilities from Rasch and 1PL models are additive and consistently ordered for item difficulties (or test-taker abilities), regardless of abilities (or items), always increasing as item easiness and test-taker ability increase. This means that the Rasch and 1PL models exhibit double cancelation and could be said to have a probabilistic relationship to conjoint measurement (Briggs et al., this volume; Embretson & Reise, 2000; Mislevy, 2018; Perline et al., 1979). Although this relationship has been the basis of attempts to establish quantitative measures with interval properties in test data, these attempts are challenging because of model–data fit limitations (see the *Scaling Perspectives* section). The relationship also reflects inconsistencies in observed item responses, the probabilistic Rasch (1960) model of latent and estimated item parameters and abilities, and the deterministic framework of observed variables in conjoint measurement (Kyngdon, 2008; Mislevy, 2018; Perline et al., 1979).

As stated in the *Scaling Perspectives* section, Rasch models are more restrictive and less likely to fit observed data than other IRT models. These model–fit limitations can prompt calls to impose statistical consequences to the data or edit or remove nonfitting responses (Wright, 1977). Examples include "reinterpreting or modifying the frame of reference" (Andrich, 1988, p. 62) or reducing the intended scale:

> The broader the domain of interest, the more difficult it will be to make targeted and testable hypotheses. This would suggest that vertical scales could only be plausibly supported for more narrowly defined latent variables. (Briggs, 2013, pp. 219–220)

Removal of test-taker data could potentially limit intended scale interpretations, such as by narrowing the intended scaling population (i.e., changing *P* in Table 11.1). Removal of nonfitting items could result in narrowing the construct defined in test specifications.

ABILITY ESTIMATES FROM 2PL AND 3PL IRT MODELS The 2PL (unique a_is and $c_i\text{s} = 0$) and 3PL (unique a_is and $c_i\text{s} > 0$) IRT models can be used to produce more complex ability estimates with unique contributions from individual items. The MLE and EAP estimates from 2PL and 3PL models reflect patterns of correct and incorrect responses to individual items, such that they differ from and convey more information than the TCC estimates. The MLE for the 2PL IRT model reflects item pattern scores and an optimally weighted test score where item scores are weighted by their discrimination, $w_i = Da_i$, and the resulting weighted test score is the sufficient statistic for θ. A TCC procedure usually based on equally weighted item scores differs from the MLE, but can produce the MLE estimate when modified to sum optimally weighted item scores (Lord, 1980). For the 3PL model, there is no sufficient statistic for θ, and the MLE estimate reflects a test score summed using optimal weights defined as $w_i = \frac{Pr(V_i = 1|\theta, a_i, b_i, c_i) - c_i}{1 - c_i} \frac{Da_i}{Pr(V_i = 1|\theta, a_i, b_i, c_i)}$.

The MLE, EAP, and TCC ability estimates based on 2PL and 3PL IRT models have unique properties (S. Kim & Moses, 2025; Kolen, 2006; Kolen et al., 2011; Thissen & Orlando, 2001; Yen & Fitzpatrick, 2006). MLEs (Equation 4) are asymptotically unbiased estimates of θ, though for tests of realistic length they have conditional biases that are positively correlated with θ (Lord, 1983) and an overall variance that is larger than that of θ. MLE estimates are unavailable for some response patterns, such as those where the responses to the test items are all incorrect or all correct. EAP estimates from Equation 5 have conditional biases that are negatively correlated with θ (Lord, 1986), and they are less variable than the MLEs. The EAP estimates reflect shrinkage to the prior distribution of θ, meaning that θ values are overestimated for test takers below the mean in the population and underestimated for test takers above the mean. θ estimates based on EAP are available for all item response patterns. Ability estimates based on EAP and MLE item pattern scores have smaller standard errors than those using the usual TCC approach, though their similarity with TCC estimates increases for longer tests with items that have smaller guessing effects (Yen & Fitzpatrick, 2006).

Ability estimates for 2PL and 3PL IRT models have implications for model–data fit, theoretical measurement properties, and score interpretations. These IRT models fit observed test data more closely than Rasch and 1PL models, but do so with added complexity. From a theoretical measurement perspective, interval properties in ability estimates from 2PL and 3PL IRT models are more difficult to defend than for Rasch models. Item characteristic curves for given items can cross for higher and lower test-taker abilities, and this nonadditivity means that 2PL and 3PL IRT models are not guaranteed to meet the double cancelation condition or have a probabilistic relationship to conjoint measurement. Ability estimates are nevertheless sometimes described as interval scales because 2PL and 3PL IRT models produce probabilities that are invariant under different linear transformations of their parameters for a specified functional form and population distribution (contrasting points are made about this description in Briggs et al., this volume, endnote 10, and Mislevy, 2018, pp. 316–317). Empirical research on interval scales from these IRT models has been encouraged (Michell, 2008), and statistical tests may be possible for the 2PL model (Kyngdon, 2011).

From an interpretational perspective, MLE and EAP ability estimates from 2PL and 3PL models are more complex and more difficult to understand than those based on simple sums of equally weighted item scores (i.e., the usual TCC). The weighting of individual items in the scoring is likely to be less closely aligned to the intended test content as determined by test developers and described in test specifications (Kolen, 2006). The complexity of ability estimates based on pattern scoring implies more complicated scoring rules and increased difficulty in communicating, understanding, and explaining scoring and advising on how test takers might maximize their performance. For MLEs based on the 2PL model, items with higher discriminations make greater contributions to test takers' ability estimates, which creates fairness issues when these differentially important items are not communicated to test takers (Dorans, 2012). For the 3PL model, MLEs reflect differentially weighted items for different test takers (i.e., for less able test takers, correct item responses tend to count less and are more likely attributed to guessing; Lord, 1980, p. 75). This characteristic was described by Mislevy (2012) as "unsettling" and tough to explain (p. 39). The fairness of pattern scoring based on the 3PL model is questionable and has been described as a challenge for establishing comparability in adaptive tests (Phillips, 2016, p. 258). A complication with EAP estimates is that they reflect not only test performance in item pattern scores, but also a test-taker ability distribution (Kolen, 2006).

ABILITY ESTIMATES FROM MORE COMPLEX MODELS Scoring approaches have been developed and considered that are even more complex than those that have been previously described. These approaches are developed based on a range of motivations. One is to obtain scores on noncognitive traits assessed with faking resistant forced-choice item formats using item parameter estimates obtained in previous administrations of those items in formats that are not faking resistant (Drasgow et al., 2012; Stark et al., 2005). Other efforts score tests composed of family-generated clone items using IRT methods like Equations 4–6 based on parameters from the item families rather than parameter estimates for the items test takers actually take (cautions provided in Drasgow et al., 2006; Harris, 2023; Luecht & Burke, 2020; van der Linden & Glas, 2010). Other efforts are being considered for measuring response processes like test-taking strategies, thought processes, and behaviors like reading, interpretation, and strategy formulation with response times, computer screen gazes, verbalizations, and other process information (Ercikan & Pellegrino, 2017). Deep-learning neural network models might be considered for automated scoring versions of subjective human scoring of test takers' writing (Zesch et al., 2023).

For complex scoring approaches, test takers' test performance measures might reflect not only their performance on the test and items they take, but also other unspecified aspects of the scoring procedures. Scoring based on item families can produce test-taker scores that reflect errors in the IRT parameter estimates of items they may not actually take. Complex scoring algorithms can be difficult to interpret and explain (Lottridge et al., 2023; Zesch et al., 2023), and the resulting test performance measures may reflect reduced accuracy and fairness, or unrepresentativeness in the data used to train the models (Broussard, 2020; Hussein et al., 2019; W. Lee & Harris, this volume). Performance measures that reflect test-taking processes can warrant validation efforts and attention to the processes different test takers use to optimize their performance (Kane & Mislevy, 2017; Wise, 2017). These approaches to representing test performance are mentioned here to illustrate recent trends for increased complexity and, likely, increased difficulty in explaining, justifying, and defending scales for the resulting performance measures.

Scaling Methods for Primary Scale Scores

The approaches to representing test performance described in the *Basic Unit of Scaling* section are usually considered inadequate for use as a testing program's reporting scale (Angoff, 1971; Kolen, 2006). Observed test scores represent test performance in ways that are specific to a test form, its items, and its measurement characteristics (difficulty, reliability, etc.). Under strong assumptions, IRT ability estimates may be regarded as theoretically independent of a test form and its items, but even if these assumptions are met in practice, the ability estimates have their own interpretational difficulties because of their similarity to standard normal variables, with means near zero, and small numbers with decimals that can be negative or positive. Consistent with the scaling discussions from earlier *Educational Measurement* volumes (Angoff, 1971; Flanagan, 1951; Kolen, 2006; Petersen et al., 1989), scale scores established from untransformed observed scores, $sc(Y) = Y$, or untransformed estimated IRT thetas, $sc(\widehat{\theta}) = \widehat{\theta}$, are not recommended.

The task in scaling is to develop a numeric transformation of a measure of test performance that assigns numbers or ordered indicators such that the scale scores,

$sc(Y)$, reflect monotonically increasing levels of achievement. For large-scale testing programs, this task of establishing scale transformations is intended to be applied not only to the scores of test form Y, but also to other alternate versions of Y developed, administered, and equated (see the *Equating* section) to Y. Another goal of scaling is that the scale scores be established to facilitate score interpretations and discourage misinterpretations. In this section, different scale score transformations are reviewed that focus on the structure, normative, shape, measurement, and content aspects of reporting scales. In each subsection, methods and intended interpretations are described.

Reporting Scale Basics

Some aspects to be determined for a testing program's reporting scale are structural, such as the scale score range and number of possible points. Recommendations are usually to establish the range of the scale scores in such a way that the scores cannot be easily interpreted as observed test score performance (Dorans, 2002; Kolen, 2006; Livingston, 2004). For example, scale score ranges such as 100–200 (Praxis) and 200–800 (SAT) are wide enough to impede unwarranted guesses about the relationship between the number of items and scale score points. The GRE scale score range (130–170) and the ACT range (1–36) are also not likely to be interpreted as simple transformations of raw scores.

Another aspect to consider in scale scores is the interval, or the number of possible scale score points reflected in the range of the scale scores. Most recommendations are to establish the scale score interval such that it represents the available information or precision of the test (Dorans, 2002; Flanagan, 1951; Kolen, 2006; Livingston, 2004). When intervals are too fine, the resulting scores could lead users to overinterpret score differences, such as on pre-1970 SAT scales of 200–800 in integer units (Livingston, 2004). When scale score intervals are more coarse than test performance (e.g., stanines), then the scale scores can result in lost information (Flanagan, 1951; Kolen, 2006) or test takers with very different test performance receiving the same scale score (Dorans et al., 2010).

Scale score intervals might be established based on simple recommendations, such as ensuring that the possible scale score points do not exceed the number of observed score points on the test (Dorans, 2002) or establishing desired confidence intervals for true scores or scale scores (Kolen, 2006). Kolen (2006) summarized two proposals for establishing scale score points that reflect desired confidence intervals for true scores for the Iowa Tests of Educational Development (ITED, 1958) and more generally (Kelley, as cited in Kolen, 2006, p. 165). These procedures are based on assuming that measurement error is normally distributed and constant across the scale, that the reliability of test performance is known or well estimated in population P, and that the scale score transformation is linear. With these assumptions, a scale score transformation that reflects true score confidence intervals could be established based on

$$6\sigma_{sc(Y),P} = 6\frac{h}{z_Y\sqrt{1-rel_{Y,P}}}, \tag{7}$$

where h is the width of the desired scale score interval, z_Y is the standard normal z-score associated with the desired γ 100% confidence interval, $rel_{Y,P}$ is the reliability of Y in population P, $\sigma_{sc(Y),P}$ is the standard deviation of $sc(Y)$ in P, and $6\sigma_{sc(Y),P}$, when rounded, provides a recommended number of distinct scale score points. For the ITED implementation in Equation 7, the scale score units are set such that adding or subtracting one scale score point establishes a 50% confidence interval for true scores (i.e., $h = 1$ and $z_Y \approx .6745$ for $\gamma = 50\%$). For Kelley's implementation of Equation 7, the scale score units are set such that adding or subtracting three scale score points establishes an approximate 68% confidence interval ($h = 3$ and $z_Y \approx 1$ for γ = 68%). From Kolen's (2006) summary of these proposals, Kelley's rule generally leads to about twice as many scale score points as the ITED rule. For both approaches, the number of recommended scale score points decreases as reliability decreases. Equation 7 can be used with any measure of test performance for which a reliability estimate is available, including an IRT θ. Once used, a linear scale score transformation is found that produces scale scores with a range of units that is consistent with Equation 7.

Another structural aspect of scale scores is the level of truncation, where recommendations are for the range of minimum and maximum reported scale scores to be narrower than the actual scale score range (i.e., the working range, Dorans, 2002). Truncation of the maximum scale scores has been recommended because it avoids interpretational difficulties such as conveying perfect test performance on test forms that differ in difficulty and have different untruncated scale scores (Livingston, 2004). Truncation also helps make the resulting scale scores more resistant to shifts in score distributions due to changes in the population or in the difficulties of the test forms (Dorans, 2002). Establishing a minimum reported scale score that is higher than what is suggested by the scale score transformation can be useful for avoiding meaningless distinctions at test performance levels lower than theoretical guessing levels where measurement is less precise (Livingston, 2004).

Finally, consider that a reporting scale can be established to specify the scale score that should correspond to one or more specific observed scores of Y. The situation of interest is one for which a scale score transformation is established such that two observed scores of test Y, y_1 and y_2, will have prespecified scale scores, $sc(y_1)$ and $sc(y_2)$:

$$sc(y) = sc(y_1) + (y - y_1)\frac{sc(y_2) - sc(y_1)}{y_2 - y_1} \tag{8}$$

This transformation has been described as a basis for a scale score range of 100–200 on Praxis (Livingston, 2004).

Normative

When establishing scale scores, a common desire is for the scale scores to represent a test-taker group of interest in terms of one or more statistics (i.e., statistics of a reference group or sample of population P taking test Y; Table 11.1). Scale score conversions might involve transformations of the observed test performance on Y to a specified scale score value. One commonly used linear transformation converts the mean and standard deviation of test performance in population P, $\mu_{Y,P}$ and $\sigma_{Y,P}$, to desired scale score values, $\mu_{sc(Y),P}$ and $\sigma_{sc(Y),P}$:

$$sc(y) = \mu_{sc(Y),P} + (y - \mu_{Y,P})\frac{\sigma_{sc(Y),P}}{\sigma_{Y,P}}. \tag{9}$$

Equation 9 can be used to establish raw-to-scale score transformations that satisfy several scaling criteria. For example, to satisfy a recommendation that the scale score mean be set at the center of the possible scale score range (Dorans, 2002), Equation 9 might be used to set the mean test performance for a test-taker group to be $sc(y)=500$ for a 200–800 range (SAT; Dorans, 2002) or 18 for a 1–36 range (ACT; Brennan, 1989). Also, Equation 9 might be used to produce a set of viable raw-to-scale score transformations that vary in how they satisfy different criteria for numbers of distinct scale score points (Equation 7) and how far the actual scale scores extend below and above a truncated range of scale scores, as well as to establish the means and standard deviations for nonlinear raw-to-scale score transformations (reviewed next).

Another example of a linear scale score transformation that reflects the normative information for a group of test takers is to establish a desired scale score standard deviation and a conversion of one particular score, y_1, to $sc(y_1)$,

$$sc(y) = sc(y_1) + (y - y_1)\frac{\sigma_{sc(Y),P}}{\sigma_{Y,P}}. \tag{10}$$

where $\sigma_{sc(Y),P}$ is the intended standard deviation of the $sc(y)s$ in population P. Linear raw-to-scale transformations such as Equations 9 and 10 establish scale scores as linearly increasing indications of test performance, retaining the shape (skewness and kurtosis) of the raw scores in the scale scores.

Nonlinear Transformations

Nonlinear transformations can be used to establish scale scores that encourage interpretations based on scale score basics (see the *Reporting Scale Basics* section), distribution shape, or measurement precision. Rounding scale scores to integers or other units can establish scale scores that convey an intended range of possible scale score points. Truncating the lowest and highest possible scale score values to intended values also helps to consistently maintain the structural aspects of scale scores. Nonlinear raw-to-scale score transformations might also be developed to satisfy other criteria. These might be subjected to an additional linear transformation like Equation 9 to reflect aspects such as the intended range and number of possible scale score points.

One nonlinear transformation involves establishing scale scores that are approximately normally distributed. The resulting scale scores can then be described as symmetric and reflective of a shape that is familiar to test users (Dorans, 2002; Petersen et al., 1989). Normally distributed scale scores can be established by first obtaining the percentile ranks or "percent at and below" for a measure of performance on test Y (Kolen & Brennan, 2014). For this discussion, consider the measure of test performance to be the raw scores obtained as the sum of I items, y_k, $k=1$ to I, where the actual scores range from 0 to y_I. A continuized version of the scores is assumed where each individual score, y_k, ranges from $y_k-.5$ to $y_k+.5$ and the entire set of continuized scores ranges from $-.5$ to $y_I+.5$. The percentile ranks for these continuized scores can be computed as

$$\begin{aligned} G_P(y_k) &= 100\{\sum_{y_j<y^*} Pr_P(y_j)+ \\ &\quad \left[y_k-\left(y^*-.5\right)\right]Pr_P(y^*)\},\ -.5\le y_k<y_I+.5 \\ &= 0,\ y_k<-.5 \\ &= 100,\ y_k\ge y_I+.5, \end{aligned} \tag{11}$$

where $Pr_P(y_k)$ is the relative frequency at score y_k for population P and where y^* is the closest integer to y_k such that $y^*-.5\le y_k<y^*+.5$. Values of the standard normal distribution are found, z_k, such that their cumulative distribution values, $\Phi(z_k)$, equal a function of the percentile ranks of the y_ks,

$$G_P(y_k)/100=\Phi(z_k)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_k} e^{-w^2/2}dw. \tag{12}$$

In Equation 12, w is a variable that assumes values ranging from $-\infty$ to z_k in the integration denoted by $\int_{-\infty}^{z_k}\ldots dw$. The resulting z_ks can be linearly transformed to have an intended mean and standard deviation (Equation 9) and rounded and truncated to have the intended range of possible scale scores. The normal transformation was used to establish recentered SAT scales in 1995 based on the 1990 reference group of graduating seniors (Dorans, 2002). Other examples provided by Kolen (2006) and Kolen and Brennan (2014) illustrate that the normal transformation is approximate and that its closeness to the normal distribution depends on the distribution of the raw scores.

Another motivation for nonlinear transformations is producing scale scores that provide a stabilized version of conditional measurement error. That is, testing standards call for reporting conditional standard errors of measurement (CSEMs) for several scores if CSEMs differ across the score range (AERA et al., 2014). For most psychometric models, the CSEMs of scores are small for the highest and lowest unscaled true scores and large for true scores in the middle of the range (Lord & Novick, 1968). Kolen (1988) proposed an arcsine transformation developed by Freeman and Tukey (1950) for use as a scale score transformation of number-correct true scores. Assuming that the errors of number-correct true scores follow a binomial, compound binomial, or IRT model (Kolen, 2006), scale scores having approximately equal CSEMs across their range can be produced as

$$g(y)=.5\left\{sin^{-1}\left[\frac{y}{I+1}\right]^{.5}+sin^{-1}\left[\frac{y+1}{I+1}\right]^{.5}\right\}. \tag{13}$$

The resulting scale scores can be reported and described as reflecting a single standard error of measurement (SEM) value across the range of true scores (Kolen et al., 1992). They can also be further transformed to have an intended mean and standard deviation (Equation 9) and rounded and truncated to have an intended range of possible scale scores. Arcsine transformations have been used to establish scales for the ACT (Brennan, 1989) and the 2016 redesigned SAT (Y. K. Kim et al., 2016).

Nonlinear transformations to achieve normal distributions or to stabilize CSEMs can have interpretational difficulties. Both types of transformations can result in extreme stretching of the highest and lowest scale scores, exaggerating scale score results and creating gaps (missing scale scores) for the highest and lowest levels of test performance. These methods can also produce scale scores with inconsistent properties, such that arcsine transformations can result in more skewness (less symmetry) in the scale score distributions. A conceptual issue with CSEM stabilization is that the scales are established for true scores and applied to observed scores.

To produce a scale score transformation that could achieve symmetry with less extreme conversions of the highest scale scores, Moses and Golub-Smith (2011) proposed a cubic transformation of raw scores,

$$sc(y) = \delta_0 + \delta_1 y^1 + \delta_2 y^2 + \delta_3 y^3, \tag{14}$$

where the δs are derived to achieve a desired mean, standard deviation, skewness, and kurtosis in the scale scores (and also to satisfy a constraint that the scale scores are monotonically increasing). This procedure can be used to approximate the normal distribution, but to varying degrees that produce less extreme transformations and fewer gaps at the highest and lowest scale score regions. This procedure was used to set the scales of the revised GRE (Golub-Smith & Moses, 2014). Another version of the cubic transformation was proposed by Moses and Kim (2017) to produce scale scores that would stabilize an inputted set of CSEMs. This procedure was used to set one of the scales of the 2016 redesigned SAT (Y. K. Kim et al., 2016). Moses and Kim (2017) also showed how cubic transformations could be established that satisfy a set of scaling criteria to varying degrees, such as by making scale score CSEMs more similar while also targeting scale score symmetry. Finally, because CSEMs are inputs in developing the scale score transformation, the cubic transformation can also target stabilizations for a limited number of scores involving consequential decisions (i.e., cut scores; Lord, 1985) or for different measures of test performance, such as CSEMs at IRT θs or at true scores and test performance measures for mixed format tests. Another situation where the flexibility of the cubic transformation might be utilized is stabilizing true score intervals for a specified range of observed scores (Lord & Novick, 1968), an application that would likely produce different scaling results than those produced to stabilize CSEMs given true scores. Finally, the cubic transformation could be extended to fit even higher moments in a targeted distribution (Headrick, 2002).

Interval Scales

Some discussions of psychological measures (Michell, 2008) and vertical scales (Briggs, 2013) have taken issue with claims that psychological and educational measures can be interpreted as quantitative and as having interval properties. A recommendation from these perspectives is that assumptions for the treatment or interpretation of variables as quantitative should be explicitly acknowledged or, better yet, should prompt empirical evaluations such as empirical tests of whether the variables meet conjoint measurement conditions (see the *Scaling Perspectives* and *Basic Unit of Scaling* sections, Michell, 2008). As suggested in the *Basic Unit of Scaling* section, most available measures of test performance are difficult to establish as quantitative scales for the above notion of "quantitative." Some apparent consequences have been described in the context of vertical scaling (Briggs, 2013), where the intent is to establish a scale where growth across grades can be measured and interpreted with respect to interval interpretations (see the *Vertical Scaling* section). Claims about vertical scales exhibiting interval properties might be made by test publishers, claims that have been discussed as inconsistent with growth rates observed to vary across lower versus higher grades or across lower versus higher parts of the scale score distribution (Briggs, 2013; Hoover, 1984). Empirical evaluations for the existence of intervals in vertical scaling results have been suggested by Briggs and Domingue (2013).

The starting point for establishing a scale with interval properties is a measure of test performance with interval properties. From the *Basic Unit of Scaling* section there are at least three possibilities:

- operationalist or pragmatic perspectives of the observed test scores as interval-based indices of performance for a specific test form
- Rasch ability estimates from test data shown to fit a Rasch model, specifically with empirically demonstrated double cancelation and conjoint measurement conditions
- ability estimates from 2PL or 3PL models, based on arguments about the specification and invariance of the probabilities from these models under different linear transformations

Following Stevens (1946), linear scale transformations (Equations 8, 9, and 10) of these test performance measures would retain the established, or argued-for, interval properties, whereas nonlinear transformations would not. Note that the linear transformation would apply to one of the three possibilities but not to all three, because the three possibilities cannot coexist because of inconsistencies in the assumed model (Rasch vs. 2PL or 3PL) and nonlinear relationships in observed test scores and IRT-based abilities.

Scales for Exams With Cut Scores

For K–12 and certification and licensure tests, the interest is in reporting performance for a limited number of performance categories rather than on the entire range of a score scale. For example, test takers might be placed into *Proficient* and *Nonproficient* categories or into levels of certification, achievement, or competence. The performance reporting for these types of tests would be oriented more to classification results and less to the properties of an entire range of scale scores. Cut scores and classifications would have increased importance, such that representing test performance relative to cut scores involves a classification problem with unique implications for measurement, CSEMs (Lord, 1985), and other properties of the classifications.

Some proposed procedures for cut scores are adaptions of what was described in the *Reporting Scale Basics* section. For tests with a long range of reported scores, Dorans (2002) recommended scale scores centered at the midpoint of the scale, with a working range that extends beyond the reporting range and with scale score units that do not exceed the number of possible raw score points. For tests with a small number of cut scores, Dorans et al. (2010) recommended that scales be established that are centered near the cut score(s), with score points not exceeding the number of raw score points, and with working ranges wider than reporting ranges, all of which would accommodate shifts in test difficulty and potential additions of new cut scores. For both types of tests, the score scale should be regarded as infrastructure likely to require repair and corrective action. Recommendations are for test assembly and scaling procedures that focus on the cut scores of interest. Also, CSEMs that are equal across the entire range of a test's scores might be less important than ensuring that the CSEMs are small near a cut score of interest (Lord, 1985).

Other procedures for establishing scale scores for tests with cut scores are content based and not data, sample, or population based as described in Table 11.1. The most well known of these are standard-setting procedures (Hambleton & Pitoniak, 2006; Ferrara et al., this volume), which begin with a statement of what competent test takers know and are able to do and then search for potential cut scores judged to reflect these statements. Various standard-setting methods can be used to ask a set of judges to work with or produce statements for what they think test takers should be able to do at a specific performance level and then identify through judgmental processes the test questions and performance that correspond to these statements. Although the basis of standard-setting procedures is primarily judgment and test content, the process is often supplemented with test performance data, such as item and test performance, and results indicative of the impact of cut score recommendations on scale score distributions.

Scales for Test Batteries and Composites

For a battery of tests covering different content areas, scales might be established using the same scaling methods with all tests in the battery, as taken by the same test-taker group. The resulting scales for each test in the battery can facilitate similar interpretations for scores in each of the tests, such that test takers' strengths and weaknesses are revealed in their scores on the battery. One of the earliest examples of battery scaling was proposed by Kelley (1914, 1923) for establishing comparable units (i.e., same means and standard deviations in the reference population) for the handwriting scoring method proposed by Thorndike (see the *Scaling Perspectives* section) and another scoring method. From a discussion of alternative proposals for establishing comparable scores from the two scoring methods of handwriting, Kelley proposed establishing comparable measures as standard scores similar to Equation 9. More recent examples of battery scaling are the recentered SAT (Dorans, 2002), where the SAT Math and Verbal scales were both established based on normal distributions with means of 500 and standard deviations of 110. The 2016 redesigned SAT also targeted similar Math and Evidence-Based Reading and Writing (ERW) Section score means and standard deviations (Kim et al., 2016). The ACT rescaling targeted means of 18 with constant SEMs (Brennan, 1989), but different scale score standard deviations. The scale scores for each of these batteries can encourage different types of interpretations, such that test batteries for normalized scales might facilitate references to percentiles of the normal distributions, those established with similar means and standard deviations might facilitate references to standard deviation units, and those established based on stabilized SEMs might encourage SEM-based score distinctions rather than distinctions based on standard deviation units (Kolen, 2006).

Composite scores might be formed based on linear combinations of the scores on each test in the battery. For example, the Total score of the 2016 redesigned SAT is the sum of the Math and Evidence-Based Reading and Writing Section scores, and the ACT Composite score is the average of the math, science, English/writing, and reading test scores. Both examples are reflections of nominally weighted scores. Other weighting procedures such as those based on effective weights and the proportional contribution of test score variances to the composites can produce different composite score scales, to the extent that the test scales have different standard deviations and the tests have different correlations with each other (Brennan, 1989; Kolen, 2006). The scaling characteristics of batteries and composites of multiple tests are affected by the scaling results for the individual tests, the scale score standard deviations of the individual tests, and the intercorrelations among the scale scores of the tests. Because scales are usually set and maintained on individual tests, the scale score characteristics for batteries and composites are more difficult to establish and maintain. The *Battery Scaling and Composites* section provides further discussion and an example of these issues.

Auxiliary Scales

After one or more primary scales are set, additional procedures and methods can be used to convey enhanced interpretations of the primary scale scores. For auxiliary scales, the goal is less about building interpretations directly into the scales themselves and more about supplementing the primary scales with additional interpretive information. Two categories of interest are based on normative information and on content.

Auxiliary Scales Based on Norms

Norms refer to the statistical information that could be provided and used to describe performance associated with different scale scores, such as the scale score mean, standard deviation, and percentile ranks (Equation 11 applied to $sc(Y)$) for one or more test-taker groups. Norms might be determined for groups sampled from particular populations. Sometimes the group used to establish norms is the same sample from population P used to establish the scale scores, in which case the norms of $sc(Y)$ are consistent with the statistics (i.e., means, standard deviations) directly established in the distribution of $sc(Y)$. Previous volumes of *Educational Measurement* have listed options for norms groups and reporting (Angoff, 1971; Flanagan, 1951; Kolen, 2006; Petersen et al., 1989), presented here with an update.

- National norms: Describe test performance from nationally representative test takers at the age or educational level for which the test was designed.
- Local norms: Describe test performance from test takers from specific educational or geographic units.
- User norms: Describe test performance from test takers who take the test during a given time period.
- Convenience norms: Describe test performance from test takers who are available at the time a test is constructed.
- Group-level norms: Describe aggregated test performance from groups, such as average performance for schools, districts, or states.
- Item-level norms: Describe the performance on specific items for a norm group.
- Skill-level norms: Describe the performance on sets of items measuring a particular skill for a norm group.
- Growth norms: Describe test performance for a group of students (e.g., K—8) that takes at least two tests at different grades, usually where separate sets of norms on the second test are reported for subgroups of students obtaining different scale scores on the earlier test (Betebenner, 2009; Castellano & Ho, 2013).

Typically, national norms that describe test performance for nationally representative test takers need the most technically elaborate norming studies. National norming studies require that the national population of interest is precisely defined in terms of students, schools, etc. A sampling plan is developed to include test takers for whom norms will produce accurate estimates of the population norms. National norming studies may use several sampling approaches to increase the representativeness of the test takers and resulting norms, including simple random sampling, random sampling within defined strata of student characteristics (e.g., geographic region, school type), cluster sampling of test-taker groups (e.g., schools), or systematic random sampling that reflects the ordering of a variable related to the test scores of interest in the norms study. Often, national norming studies use a combination of sampling and statistical weighting procedures to obtain schools and test takers and increase their national representativeness. Approaches to national norming studies by the National Assessment of Educational Progress (NAEP, https://nces.ed.gov) and by ACT were compared in terms of their methods, complexity, and practical challenges by Kolen (2006). Other types of norms (user, convenience, etc.) might be produced using less elaborate procedures (e.g., test takers available at the time the norms are produced). Such norming groups may be less stable and more likely to change (Petersen et al., 1989).

Auxiliary Scales Based on Content

Some auxiliary scales are attempts to supplement primary score scales with content-based descriptions of specific scale scores. One example involves item mapping, or finding a set of items that represents various scale score points. In the case of NAEP, these items "help to illustrate what students know and can do in NAEP subject areas by positioning descriptions of individual assessment items along the NAEP scale at each grade level" (https://nces.ed.gov). For item mapping, the scores of individual items are regressed on scale scores using logistic regression or an IRT model. Then a response probability is adopted to divide test takers into groups with "higher" and "lower" probabilities of success on test items. The actual values used to divide the higher and lower probability groups could be .5/.5, .8/.2, or other probabilities for dichotomously scored items. In NAEP, these groups are divided using probabilities based on 74% for MC items and 65% for CR items. These response probability values were obtained based on work by Huynh (1998), who showed that these values correspond to the maximum IRT-based information provided by a correct response to MC or CR items. The result of item mapping is a list of test questions that represents various scale score points, with content-based descriptions that take the form of phrases that describe what test takers can do correctly, in general terms rather than for one or more specific items (Kolen, 2006; Kolen & Brennan, 2014).

Scale anchoring is another type of content-based auxiliary scale, with the goal of providing general statements about what students obtaining different scale scores would know and are able to do (Kolen, 2006). Item maps are a first step to scale anchoring. Additional steps involve choosing a set of scale score points across the score scale range, where items that map at or near these points are chosen to represent these points. Then subject matter experts review the items mapping near each point and produce statements presenting the skills of test takers scoring at these points, assuming test takers also know and are able to do all of the skills in the statements at or below the given score level. Scale anchoring has been used in NAEP (Allen et al., 1999) and in ACT (2001). The Binet–Simon (Binet & Simon, 1916) test described in the *Scaling Perspectives* section can also be viewed as a test with sets of items specifically anchored to different scale scores. Item-mapping and scale-anchoring procedures have raised questions about whether their outcomes actually facilitate score interpretations (Forsyth, 1991). The content-based labeling produced with scale anchoring appears to be the result of overly subjective and confusing judgmental processes (Pellegrino et al., 1999) that may not be supported by statistical analyses (Haberman, Sinharay, & Lee, 2011; Thurstone, 1925) or by items that are sufficiently discriminating (Dorans, 2018).

Summary of Primary and Auxiliary Scales

Reporting scales should be viewed as infrastructure in need of monitoring and possible repair (Dorans, 2002). For the primary scaling approaches reviewed in the *Scaling* section, potentially important checks for primary scales include checks for excessive accumulations at the top, middle, or bottom parts of the scale; checks that the score means do not excessively depart from the midpoint of the scale; and other checks that properties like distribution shapes, CSEM values, etc., are established and maintained as intended. As described in the *Scales for Exams with Cut Scores* section, primary scales can also be established and monitored in terms of cut scores of interest, including distributions resulting from pass/fail decisions and CSEMs at or near those cut scores. These possible checks suggest different types of monitoring for primary scales established in different testing contexts. In admissions testing where testing companies develop tests for users to employ different parts of a relatively long scale, monitoring should focus on the entire score range. In K–12 and certification and licensure programs, the focus is primarily on classifications at one or very few cut points, and the monitoring of scale scores would likely focus on pass/fail distributions, classification accuracies, and CSEMs at specific cut scores. Scale maintenance for large-scale survey assessments like NAEP might focus on the estimation of population and subpopulation scale score distributions and classifications over time. In each of these contexts, scales established in special studies should be checked to ensure that their results will accurately apply beyond these scaling studies and to the intended testing population. The auxiliary scales that might be established and reported for a particular testing program should also be monitored for potential changes due to testing population changes and other potential changes in content-based descriptions. The equating procedures described in the *Equating* section are needed for preserving established primary and auxiliary scales.

EQUATING

For a large-scale testing program, the establishment of a reporting scale as described in the *Scaling* section usually comprises only an initial step for a single test form (Y). Usually Y is the first of many test forms to be developed and administered. Additional work is needed to develop, administer, and report scores for the subsequently developed test forms for that testing program. Equating is used to address unintended difficulty differences in the scores of these alternate forms, so that the previously established reporting scale is maintained (Holland & Dorans, 2006; Kolen & Brennan, 2014). This section describes equating in an updated version of the discussion from Holland and Dorans (2006), which itself was a continuation of previous discussions from Petersen et al. (1989), Angoff (1971), Flanagan (1951), and others.

The *Distinguishing Equating From Other Forms of Linking* section describes some general concepts about equating and its distinctions from other types of linking based on goals, interpretations, and computational procedures. Equating history and requirements are covered in the *History of Equating* and *Equating Requirements* sections. The data collection designs used in different types of equating are described in the *Data Collection Designs* section. The final section covers *Methodological Implementations of Equating*. The different equating approaches covered in these discussions are summarized in Table 11.2 in terms of the equating definition, method(s), assumptions, and likely data collection designs.

Table 11.2 Summary of Types of Equating

Definition of Equating	Method(s)	Assumptions	Data Collection Design(s)
Align the means and standard deviations of observed scores of alternate test forms (common population)	Linear scale alignment	Tests measure the same construct; population invariance; equity	Randomly equivalent groups; single groups; counterbalanced
Align the distributions of observed scores of alternate test forms (common population)	Equipercentile		
Align the means, standard deviations, and/or distributions of observed scores of alternate test forms (different populations)	Tucker, Braun–Holland; frequency estimation/poststratification	Tests measure the same construct; population invariance in the conditional test-given-anchor distributions; equity	NEAT
	Chained linear; chained equipercentile	Tests measure the same construct; population invariance in the two chained functions; equity	NEAT
	Levine observed score; modified frequency estimation	Tests and anchors are congeneric; population invariance in the conditional Test-given- $\hat{\tau}_A$ distributions; equity	NEAT
	IRT observed score	Items measure the same construct; the IRT model fits the data; IRT parameter estimates for all items are on the same scale and invariant across administrations; equity	NEAT (this method is also used with randomly equivalent groups, single group, and counterbalanced)
Align the means and standard deviations of true scores of alternate test forms (different populations)	Levine true score	Tests and anchors are congeneric; true score results can be applied to observed scores; equity	NEAT
Link true scores via an expected test-given-ability relationship based on an IRT model	IRT true score	Items measure the same construct; the IRT model fits the data; IRT parameter estimates for all items are on the same scale and invariant across administrations; true score results can be applied to observed scores; equity	Randomly equivalent groups; single group; counterbalanced; NEAT
Score test takers using item parameter estimates obtained and linked in previous administrations	IRT ability estimation	Items measure the same construct; the IRT model fits the data from the previous calibrations; parameter estimates are invariant across administrations with minimal context and order effects; equity	Pre-equating

Note. IRT = item response theory; NEAT = nonequivalent groups with anchor test.

Distinguishing Equating From Other Forms of Linking

Equating is one type of linking procedure for transforming the scores of one test form to another (Holland, 2007; Holland & Dorans, 2006; Pommerich, 2016). Specifically, equating adjusts the scores of test forms constructed and administered in the same way for unintended difficulty differences. Within a linking hierarchy, equating is the strongest among several types of linking, distinguished in terms of interpretations and computational methods (Angoff, 1971; Flanagan, 1951; Holland & Dorans, 2006; Mislevy, 1992; Petersen et al., 1989).

Different categories of linking can be used to achieve different goals and support particular interpretations of the linked scores. When equatings of alternate forms of the same test are produced (X and Y), the expected result is interchangeability in the reported scores, meaning that when form X is equated to form Y, the resulting X-to-Y scores can be used in place of Y's scores for any purpose, as if those scores came from the same test (Dorans, 2013; Pommerich, 2016). For other types of linking involving different tests (i.e., X and Y are different tests), weaker and more limited results are expected. One alternative goal is comparable scales (Angoff, 1971; Kelley, 1923; Pommerich, 2016), which is one component of interchangeability that can be defined as the result of a symmetric alignment of scale score distributions that facilitates comparisons of the resulting scores. In the absence of interchangeability, comparability is group dependent (see the *Linking* section).

Another goal of linking is "best" prediction, which is an asymmetric conversion from the scores and scales of one test to another, such that prediction error is minimized. Predictions are distinguished from interchangeability and comparability in that predictions are often group dependent and always asymmetric (i.e., for given X scores, the best prediction from X to Y does not equal the best prediction from Y to X).

Scale-Aligning Computations

Scale-aligning methods are one type of computation used in test linking, including equating (Holland & Dorans, 2006). These methods can be described as score transformations that result in the scores of two test forms having the same distribution (Kelley, 1923). For example, consider test Y as a form for which a scale score transformation has been established, $sc(Y)$, and X is another test form whose scores are intended to be transformed to align to the same scale. The transformation of X to the scale of Y may be achieved through matching the marginal distributions of these forms obtained from test takers representing target population T (Holland & Dorans, 2006, p. 202, defined T as the source of the equating data and also presented other perspectives on target populations in equating). An equipercentile X-to-Y transformation produces Y scores with percentile ranks that equal those of every X score for X and Y test takers from the target population T. The equipercentile linking is produced through computing percentile ranks at each score of X (Equation 11 for score x on the distribution of X, $F_T(x)$) and obtaining the inverse of Equation 11 on the distribution of Y (i.e., $G_T^{-1}[y]$) in T,

$$e_{Y,T}(x) = G_T^{-1}[F_T(x)]. \tag{15}$$

Equation 15 produces a symmetric X-to-Yconversion that equals the inverse of the Y -to-X conversion at the same x scores, $e_{X,T}^{-1}(y)$ at x. The X-to-Y scores have a distribution on X that approximately equals the distribution of the Y scores in population T. Simpler versions of Equation 15 can also be produced by using a version of Equation 9 to solve for X-to-Y scores that match the mean and standard deviation of Y but not other aspects of Y's distribution,

$$l_{Y,T}(x) = \mu_{Y,T} + (x - \mu_{X,T})\frac{\sigma_{Y,T}}{\sigma_{X,T}}, \tag{16}$$

where $\mu_{X,T}$ and $\mu_{X,T}$ denote the population means of X and Y in T, and $\sigma_{X,T}$ and $\sigma_{Y,T}$ denote the population standard deviations. Equations 15 and 16 establish comparability through symmetric transformations that match the score distributions (Angoff, 1971; Holland & Dorans, 2006; Kelley, 1923; Pommerich, 2016). Once the results are obtained from either Equation 15 or Equation 16, X-to-Y scores can be transformed to the reporting scale by applying one of the $sc(Y)$ scale transformations described in the *Scaling* section.

Projecting and Predicting Computations

Projecting and predicting are types of linking procedures that differ from scale-aligning approaches like equating. Rather than aligning the scores of form X and form Y to have comparable score distributions for a particular group of test takers, these methods are used for the goal of producing an asymmetric, best prediction of scores or a projection of score distributions for one test to those of another. Predictions and projections can also be produced for the IRT-based thetas from different tests (Holland & Hoskens, 2003; Thissen et al., 2015). One example of a projected distribution is the percentile rank of $Y = y$ for a subgroup of test takers obtaining a specific score on $X(= x)$, $F_T(y|x)$. Another example is the prediction of the means of Y given X using a linear X-to-Y regression,

$$reg_{Y,T}(x) = \mu_{Y,T} + (x - \mu_{X,T})\frac{\sigma_{Y,T}}{\sigma_{X,T}}\rho_{XY,T}, \tag{17}$$

where $\rho_{XY,T}$ is the XY correlation. The prediction shown in Equation 17 is a linear regression that provides an expectation of Y given X that minimizes prediction error, the variance of $Y - reg_{Y,T}(x)$, an essential supplement to Equation 17's point estimates (Thissen et al., 2015). Unlike some computational methods for scale aligning, projecting and predicting functions require data collection designs where test takers take both X and Y. Also unlike scale-aligning functions, predicting functions are not symmetric, meaning that $reg_{X,T}^{-1}(y)$ at x is not equal to $reg_{Y,T}(x)$.

Some results are of special interest for situations in which there is a perfect one-to-one relationship in X and Y. For linear functions, this situation means that the correlation between X and Y is 1 and the X-to-Y regression in Equation 17 produces the same results as the X-to-Y scale alignment in Equation 16. This result has several uses. One use is defining a situation in which a regression can be used to align scales. For example, consider the second half of Equation 6, $\sum_i w_i Pr\,(V_i = 1 | \theta, a_i, b_i, c_i) = \tau_Y(\theta)$, which can be described as a regression of IRT-implied true scores on θ (Lord & Novick, 1968, p. 386), such that these true scores have a perfect and functional (not statistical) nonlinear relationship with θ under the IRT model. When same-scale IRT parameter estimates are available for the items of tests X and Y, the IRT-based true score regressions of X and Y can be combined to produce a scale alignment of the true scores of X and Y, $\tau_Y\left[\tau_X^{-1}(\theta)\right]$, where $\tau_X^{-1}(\theta)$ denotes the inverse of the θ-to-τ_X function (i.e., τ_X-to-θ, Lord, 1980). In addition to these IRT applications, perfectly correlated X and Y true scores might also be assumed and used to produce scale alignments of τ_X and τ_Y based on classical congeneric theory (see the *Equating Requirements* section; Kolen & Brennan, 2014; Levine, 1955). Another use of the relationship of Equations 16 and 17 is to describe the strength of a linking of tests with scores that are not perfectly correlated. For example, in concordances of different tests (see the *Concordances* section), the correlation has been used as a measure to diagnose the strength of the resulting concordances (see the *Correlations and Prediction Error* section), such as by describing the extent to which a scale alignment might be used to produce score predictions that approximate the accuracy of projection or prediction functions (Dorans, 1999; Moses, 2014a).

History of Equating

The origins of equating date to the early 20th century, when an intelligence test known as the Army Alpha was developed for selection into the military for World War I (Yoakum & Yerkes, 1920). Alternate test forms were developed by randomly assigning items to the forms, to prevent cheating and to establish forms of approximately equal difficulty. The difficulties of the alternate forms were subjected to empirical evaluations, and for the subset of forms surviving these evaluations, the unadjusted observed scores were treated as interchangeable. The linear linking function (Equation 16) was recommended over the nonsymmetric regression function (Equation 17) for establishing comparable measures on the basis of approximating scale score distributions (Otis, 1922; Thorndike, 1922). Kelley (1923) provided more discussions of linear linking functions and an introduction of the equipercentile method (Equation 15). Nevertheless, “the need, or at least the desire, to equate scores on alternate forms of the same test probably arose decades after the invention of scaling methods and of the two standard methods for equating—the linear and equipercentile methods” (Holland & Dorans, 2006, p. 196).

Equating was described in general and in practice for the Cooperative Achievement Tests by Flanagan (1939, 1951). These discussions covered equating for the data collection designs currently referred to as the single group, randomly equivalent groups, and counterbalanced designs (see the *Data Collection Designs* section). Other emphases were on form construction and linear and equipercentile methods. Test equating for nonequivalent groups with an anchor test (the NEAT design, see the *Data Collection Designs* section) has been traced to the SAT, which administered two editions in 1938 and employed anchor equating in 1941. For this SAT application, a method attributed to Tucker and eventually named Tucker equating (Angoff, 1971) was used to implement a linear scale alignment (Equation 16) with regression-based predictions of the X and Y means and standard deviations for a hypothetical target group (Gulliksen, 1950; Holland & Dorans, 2006; Lord, 1950).

Since the early SAT equatings, additional discussions and uses have been provided, including Flanagan's (1951) comprehensive treatment of test development methods in equating, computational methods, data collection designs, a definition of comparability in equating, and, notably, a warning about likely population sensitivities and group dependences in linking results (p. 748). Lord (1950) provided a statistical overview of linear equating procedures and standard errors for various data collection designs. Angoff (1971) provided another comprehensive discussion of equating, computational methods and data collection designs, which, in contrast with Flanagan (1951), emphasized the goal of equating to produce results that are relatively group invariant (p. 563). Angoff also described situations where a linear equating function approximates the equipercentile function (i.e., when the shapes of the X and Y distributions are similar) and recommended linear functions for these situations (p. 564). Angoff's (1971) chapter has been a foundational and influential reference for decades of practice and literature on equating, scaling, and linking.

Updates to the statistical aspects of equating were provided by Holland and Rubin (1982), including a discussion of the mathematical properties of commonly used equating methods (Braun & Holland, 1982) and section pre-equating (Holland & Wightman, 1982). Interspersed in this history, proposals were made to describe and implement equating based on measurement theory, including classical true score theory (Levine, 1955), IRT (Lord, 1980; Lord & Wingersky, 1984), and other works focusing on theoretical properties such as equity of equating results with respect to true scores (Hanson, 1991; Morris, 1982) and IRT (van der Linden, 2011). The IRT developments were especially important for providing alternative theoretical and practical implementations for equating, including proposals for item-level pre-equating and IRT-based adaptive testing (Lord, 1980). Discussions and updates have continued, including the chapters in the third and fourth editions of *Educational Measurement* (Holland & Dorans, 2006; Petersen et al., 1989), subpopulation invariance evaluations (Angoff & Cowell, 1986; Dorans & Holland, 2000; Harris & Kolen, 1986), several texts (Dorans et al., 2007; Holland & Dorans, 2006; Kolen & Brennan, 2014; Linn, 1993; Livingston, 2004; Mislevy, 1992; von Davier, 2011; von Davier et al., 2004a), and software packages (Brennan, 2004; Brennan et al., 2009; Gonzalez & Wiberg, 2017).

In practice, equating has been implemented using different data collection designs for different testing programs, including programs that equate with randomly equivalent groups designs (ACT and the 2016 redesigned SAT) and others that use anchor tests to equate through nonequivalent administration groups designs (the SAT prior to the 2016 redesign). Other programs use designs suited for small samples (Puhan et al., 2009), and for pre-equating (CLEP or College-Level Examination Program, Gao et al., 2012). These implementations of test equating are summarized in Table 11.2.

Equating Requirements

In discussions of linking frameworks (Holland & Dorans, 2006), equating is described as the strongest form of linking and is subject to the strictest requirements. These requirements were listed by Holland and Dorans (2006) as a culmination of several previous discussions of equating (Angoff, 1971; Dorans & Holland, 2000; Flanagan, 1951; Kolen & Brennan, 2014; Linn, 1993; Mislevy, 1992; Petersen et al., 1989). The equating requirements are summarized in this section, including updates and clarifications from Brennan (2010), Dorans and Walker (2007), and Kolen (2007).

An equating can be used to align the scales of X and Y, such that X-to-Y and Y scores can be used interchangeably, thereby maintaining a testing program's reporting scale. For an X-to-Y transformation to be considered an equating, the X and Y test forms and the X-to-Y transformation must satisfy the following requirements:

a. The equal construct requirement: Test forms X and Y should measure the same constructs.
b. The equal conditions of measurement requirement: The same administration and measurement conditions should be used for X and Y.
c. The equal and high reliability requirement: The reliabilities of X and Y, $rel_{X,T}$ and $rel_{Y,T}$, should be equal and high.
d. The equity requirement: The X-to-Y equating should make it a matter of indifference whether test takers take X or Y, for test takers at every given ability level.
e. The (sub)population invariance requirement: The X-to-Y equating should be population invariant, in that the choice of (sub)population used to estimate it does not matter.
f. The symmetry requirement: The X-to-Y equating function should be the inverse of the Y-to-X equating function.

Discussions of the equating requirements describe them as aspirational concepts (Brennan, 2010; Dorans & Moses, 2023) that provide an intuitive theory of test equating (Holland & Dorans, 2006). These requirements are not likely to be perfectly met in practice, sometimes prompting criticisms for being vague, impractical, stringent, unnecessary or impossible (Dorans & Holland, 2000; Livingston, 2004; Lord, 1980). Nevertheless, the equating requirements specify the goals of the test development, administration, scoring, and linking processes that are useful for producing interchangeable scores that maintain the reporting scales of a testing program for a target population of test takers. As described in *Linking*, the equating requirements are also useful for characterizing nonequating types of linking.

Equating requires that the test forms being equated are assembled according to the same content and statistical specifications, such that they exhibit the same measurement properties (Requirement a). Equating Requirement b emphasizes consistency in administration conditions for tests that are equated (Kolen, 2007). In describing conditions of measurement, Kolen (2007) listed aspects under the control of the test developer such as that tests X and Y are administered with the same instructions, layout, timing, scoring procedures, aids, and modes (computer or paper–pencil). Other measurement conditions that are not under the direct control of the test developer can also affect the quality of equating outcomes, including stakes for test performance, reasons test takers take the test, and type of test preparation activities.

Requirement c, that the test forms are highly and equally reliable, has received ongoing and updated clarifications. The equal reliability aspect of this requirement reflects earlier recommendations, meaning that the test forms being equated must have the same measurement precision. The requirement that forms' reliabilities be high is a more recent addition, which conveys that equating requires scores that are precise and close reflections of their true scores, that support high correlations and prediction accuracy, and that make subpopulation invariance (Requirement e) and equity (Requirement d) more likely (Brennan, 2010; Dorans & Walker, 2007; Flanagan, 1951; Holland & Hoskens, 2003; Kolen, 2004; Lord, 1980).

As stated in the *Scaling* section, many issues concerning reliability, true score variance, and error variance are, at best, imperfectly understood. They are also increasingly complex when reliability and equating are considered in greater detail and in relation to other equating requirements. The reliability requirement depends on consistently following test development procedures and specifications for test length, content, items, administration, and scoring procedures (Requirements a and b). Test forms developed from different sets of items that meet the same specifications could be assumed to be parallel in the classical test theory sense (equal reliabilities, equal true scores, and errors with equal variances). In practice, assembled forms can be imperfectly or nominally parallel (Lord & Novick, 1968) and might have unintended differences, such as differences in their mean scores attributable to their unique samples of items. Equating can be described as the application of scale-aligning adjustments that address errors attributable to forms' item samples, errors that are realized in fixed difficulty differences of the specific form(s) being equated (Li, 2023; Moses & Kim, 2015). Equating does not address the reduction of relative errors (i.e., interactions of test takers and items, as well as other unspecified sources of error; Li, 2023).[1]

Forms' error variances can also be considered conditional on test-taker ability, variances that are assumed to be equal according to Equity Requirement d. One proposal based on equity applies conditional, ability-specific scale-aligning adjustments (van der Linden, 2011, 2013), which could be interpreted as an attempt to adjust scores for IRT-implied relative errors, with the forms' items and parameters being treated as fixed. As described in the *Local, Ability-Specific Test Linking* section, these local adjustments likely violate other equating requirements concerning consistency in scoring rules for a given form (Requirement b) and subpopulation invariance (Requirement e).

Different score models and associated reliability measures have equating implications and provide unique representations of reliability and error, including model-implied error variances based on fixed items (commonly used IRT models), score models that treat items as random samples, and more complex score models that account for errors across a range of admissible measurement conditions. These issues illustrate the lack of a single framework for measurement error and other types of equating error, the need for greater specificity in reliability and equating, and the importance of test form reliabilities that are high and equal in some sense for reducing the impact of these ambiguities on equating results.

Equity Requirement d states that the conditional distributions of Y and X-to-Y equated scores are equal for test takers at specific values of a latent ability based on a particular measurement model (Brennan, 2010; Hanson, 1991; Lord, 1980; Morris, 1982; van der Linden, 2011). Equity was originally defined in terms of IRT models, θs, and θ-conditional frequency distributions (Lord, 1980). Equity is commonly expressed based on θ-conditional cumulative distributions (Kolen & Brennan, 2014; van der Linden, 2011),

$$F_\theta\left[e_{Y,T}(x)\right] = G_\theta[y] \quad \text{for all } \theta, \tag{18}$$

where the θ subscript indicates conditioning on θ.

Equity discussions based on classical test theory have focused on relationships among equating functions of observed scores, true scores, and equating requirements (Holland & Dorans, 2006; Hanson, 1991; Kolen & Brennan, 2014; Morris, 1982). Holland and Dorans (2006) attributed an equity theorem to Hanson (1991) and synthesized several equating requirements. To summarize, assume that X and Y measure the same construct and follow a congeneric model (i.e., their true scores are perfectly and linearly related) and the functional relationship of the true scores, τ_X and τ_Y, can be expressed as a scale aligning function,

$$\tau_Y = l_{\tau_Y}(\tau_X) = \mu_{Y,T} + (\tau_X - \mu_{X,T})\frac{\sigma_{Y,T}}{\sigma_{X,T}}\frac{\sqrt{rel_{Y,T}}}{\sqrt{rel_{X,T}}}. \tag{19}$$

Substituting observed X scores into true score conversions like Equation 19, $l_{\tau_Y}(x)$ satisfies first-order equity where the conditional mean of $l_{\tau_Y}(X)$ given $l_{\tau_Y}(\tau_X) = \tau_Y$ equals the conditional mean of Y given τ_Y for all τs. Hanson (1991) showed this first-order equity result in Levine true score equating for observed scores (see the *True Scores* section) with the nonequivalent groups with anchor test (NEAT) design (see the *Data Collection Designs* section). Holland and Dorans (2006) additionally stated that when X and Y are equally reliable, $l_{\tau_Y}(x)$ equals $l_{Y,T}(x)$ in Equation 16 and satisfies second-order equity where the conditional variance of $l_{\tau_Y}(X)$ given $l_{\tau_Y}(\tau_X) = \tau_Y$ equals the conditional variance of Y given τ_Y for all τs. Brennan (2010) provided a related discussion that considered quadratically represented curvilinear equating functions, showing that curvilinear equating functions are more likely to satisfy first- and second-order equity if the reliabilities of X and Y are not only equal, but also high (nearly linear equating functions are also helpful, but less realistic). Altogether, these results connect several equating requirements, such that X and Y are assumed to measure the same constructs (Requirement a), are administered under the same conditions (Requirement b), with scores that have equal and high reliabilities (Requirement c), and where the equating is a symmetric scale alignment of the forms' observed scores (Requirement f) that approximates the functional relationship of their true scores and achieves first- and second-order equity (Requirement d) (Holland & Dorans, 2006). The results represent one of the most general theories of equating available, including all but one equating requirement (i.e., the subpopulation invariance requirement is out of scope, since the results are based on a fixed target population, T; Holland & Dorans, 2006).

Equating Requirement e emphasizes that the same X-to-Y linking results should be obtained if estimated in T and in subpopulation T_g,

$$e_{Y,Tg}(x) = e_{Y,T}(x) \quad \text{for each } T_g. \tag{20}$$

The subpopulation invariance requirement is more likely met when X and Y are constructed to be similar and of high reliability (Dorans & Holland, 2000; Flanagan, 1951; Kolen, 2004). The invariance requirement of equating has been especially important for use in the empirical evaluations of nonequating types of linking (see the *Linking* section) and also for linkings intended to be used as equatings. Indices for quantifying the extent to which equating functions vary by subgroup are summarized in the *Subpopulation Invariance Evaluations* section.

Requirement f indicates that asymmetric prediction and projection functions cannot be used as equating functions because asymmetric functions mean that it is not a matter of indifference which form test takers take. Symmetric functions like Equations 15 and 16 can be described as a computational realization of comparable scales (i.e., comparable scale score distributions; Holland & Dorans, 2006; Kelley, 1923).

Data Collection Designs

Data collection designs for test equating are summarized in Table 11.1 and described in detail in several texts (Angoff, 1971; Dorans, 2018; Dorans et al., 2010; Holland & Dorans, 2006; Kolen & Brennan, 2014; Kolen, 2007; Petersen et al., 1989; von Davier et al., 2004a). Each design is an approach to obtaining test data such that an equating of the forms' scores will estimate differences in test difficulty and other characteristics, where the estimates (a) control for the abilities of the test-taker groups taking both X and $Y$and (b) reflect target population T. Successful data collection designs are essential for obtaining data from which accurate equating and other scale alignment and projection functions can be estimated.

The simplest test equating design shown in Table 11.1 is the single group design, where one sample of test takers from population $P\,(=T)$ takes both X and Y. This design can use relatively small samples to produce statistics that accurately reflect their population values. The necessary conditions for this design are that it must be realistic for test takers to take both tests and that the resulting data are not adversely affected by timing issues, testing fatigue, learning, or order effects. Scale-aligning, predicting, and projecting methods are all possible with the single group design, but satisfying the necessary conditions can be challenging. One advantage of this design is that a correlation between the two tests can be estimated to enable a direct check on Requirements a and c, provided that Requirement b is also met.

Another design shown in Table 11.1 is the randomly equivalent groups design, where two independent samples of test takers from population $P\,(=T)$ are randomly assigned to take X or Y. In large-scale testing, the random assignment is usually implemented with spiraling, where X and Y are alternated in their delivery, resulting in systematically assigned test forms to the test-taker groups. Groups end up being more equivalent than would be expected with simple random sampling (von Davier et al., 2004a). Because independent samples are used, the randomly equivalent groups design requires larger sample sizes to produce statistics for test forms X and Y with the same precision as obtained with the single group design. When test-taker sample sizes are large and the test forms can be reused and readministered without security problems, the randomly equivalent groups design has advantages in that it avoids order effects that can arise with the single group design. Scale-aligning methods are used with the randomly equivalent groups design.

In the counterbalanced design shown in Table 11.1, independent samples are drawn from population $P\,(=T)$, one where test takers take form X and then Y, and another where test takers take Y and then X. This implementation combines the single group and randomly equivalent groups designs, allowing for order effects to be estimated based on the differences in test linking results for data reflecting each testing order. To the extent that order effects are observed, various choices can be made for linking forms X and Y with the counterbalanced data (von Davier et al., 2004a), including the use of all available data to represent both testing orders, the use of only subsets of the data that do not reflect order effects (where some test takers take X first and other test takers take

Y first), and other weighted combinations of these. Usually a counterbalanced design is implemented in a special study where the estimation of order effects is of interest, rather than in a typical test administration where testing orders are fixed. As with the single group design, Requirements a and c can be checked, as can Requirement b.

Another data collection design shown in Table 11.1 is less direct, where X is linked to Y through scores on an anchor test, A. This is the NEAT design (Holland & Dorans, 2006), which is also referred to as the common-item nonequivalent groups design (Kolen & Brennan, 2014). Samples from two different populations take X or Y, and both samples take A. Because the samples represent different populations, estimates of population differences are needed for the X-to-Y linking, and the groups' performance on A provides these estimates. The NEAT design contains two single group designs. In practice, A can be a set of items common to X and Y (an internal anchor) or a test or set of items that is separate from X and Y (an external anchor). The quality of the final linking results based on internal or external anchors reflects the extent to which A is representative of X and Y. Accordingly, data collections and procedures are encouraged that strengthen the anchor test (Dorans et al., 2011; Holland & Dorans, 2006).

The NEAT design is more complex than the single group, randomly equivalent groups, and counterbalanced designs. Specifically, the NEAT design requires invariance assumptions for estimating test-taker group performance on the test that test takers do not take (assumptions about the unobserved Sample 1 from population P on Y and about the unobserved Sample 2 from population Q on X). Invariance assumptions must be made to produce an equating that applies to the target population. Two major approaches and their associated assumptions involve chaining through A and estimating X and Y distributions by projecting from A for a synthetic population defined as a combination of P and Q (described in the *Approaches to NEAT Equating* section).

An approach similar to the NEAT design involves assembling X using items from an item pool that have IRT parameter estimates on the same scale, X_A, along with other newly administered items, X_{New}. In this common item equating to a calibrated pool design (Table 11.1, Kolen & Brennan, 2014), the X_A items are used as an anchor to obtain scaled IRT parameter estimates for the X_{New} items that expand the item pool. Then, as described in the *IRT* and *Distinguishing Equating From Other Forms of Linking* sections, an IRT conversion to the θ scale, to Y, or to any other form that might be assembled from the item pool, is produced for the scores of the complete X form (X_A and X_{New}). Parameter estimates for the item pool and the equating results should all apply to population $P\ (=T)$. Common item equating to a calibrated pool is more flexible than the NEAT designs with respect to the source of the anchor items, but it requires an IRT implementation and the meeting of IRT assumptions (Kolen & Brennan, 2014).

The final rows in Table 11.1 present two pre-equating designs for linking X to Y using data from previously administered sections or individual items to equate X before it is ever administered as an intact test. Section pre-equating involves the administration of a complete old form, Y, composed of two mutually exclusive sections (Y_1 and Y_2) and one of two mutually exclusive sections of new form X (X_1 or X_2) to randomly equivalent samples of population $P\,(=T)$. The statistics for the complete X form could be estimated using data from the administration of the X_1 or X_2 sections, and these estimated statistics could be used to link X to Y (Holland & Wightman, 1982). The use of two sections is the simplest case for this design. In their discussion of section pre-equating, Petersen et al. (1989) presented implementations with three and four sections of X and Y administered to three and six random samples of P.

Pre-equating can also be implemented in individual items rather than test sections. The final row of Table 11.1 presents an implementation of item pre-equating that is similar to common item equating to an item pool except that the new items, *New*, do not contribute to the equating and scoring of X (Kolen & Brennan, 2014). That is, X is assembled from items from a calibrated item pool rather than one specific form and is pre-equated and scored using the IRT parameter estimates from these calibrated items prior to the administration of X. Then, in a subsequent step after the administration of X, scaled IRT parameter estimates are obtained for the *New* items using X and these are added to the item pool for assembling and pre-equating additional forms. Parameter estimates for the item pool and the pre-equating results for other forms assembled from the item pool should all apply to population $P\,(=T)$. A more complex item pre-equating design involving multiple tests and populations is presented by Petersen et al. (1989), and is a basis of adaptive testing and linking (see the *Linking Adaptive Tests* section).

Pre-equating designs are relatively complex and require stronger assumptions than the other designs described, namely, that the test, section, and item statistics estimated in a pre-equated administration are accurate when used to produce equating and other linking functions. Inaccuracies can result from population differences or from context and order effects where the pre-equating administration involves administering the items and subsections of X in different orders and contexts than those from the actual administration of the intact form X (Davey & Lee, 2011). In addition, for implementations where a typical IRT model is used that assumes test-taker ability is unidimensional, pre-equating could introduce biases when pre-equated IRT statistics based on unidimensionality assumptions are used to approximate a linking for the full multidimensional test (Kolen & Brennan, 2014). Estimates of X from pre-equating could be biased reflections of the actual X, resulting in inaccurate equating results that might be corrected with “postequating” using actual administration data and one of the traditional designs in Table 11.1.

Methodological Implementations of Equating

This section provides a summary of research, discussions, and studies on the methodological issues in equating. The methodological issues covered here pertain to material covered in the following sections: *Data Use and Smoothing, Continuization, Choices Among Data Collection Designs, Approaches to NEAT Equating, True Scores, IRT*, and *Accuracy Evaluations*. Many of these procedures can and should be implemented using open-source computational programs that provide critical supports for transparency and for the independent replication of results (Brennan, 2004; Brennan et al., 2009; Gonzalez & Wiberg, 2017). In addition to the methodology covered in this section, important aspects of equating implementations include checks that test that development and administration practices are standardized and similar, that quality control procedures are implemented to ensure administration conditions are followed, and that tests, items, and answer keys are all working and displayed as intended. The implications of altered questions and answer keys are covered by Kolen and Brennan (2014, pp. 331–336).

Data Use and Smoothing

Equating procedures are statistical operations performed on sample data. Ideally, the sample data should come from an established data collection design and should be a large and sufficiently representative sample that can produce precise adjustments for the test linking. The question How large? depends on the data collection design and can be addressed based on standard errors of equating for the specific design (Kolen & Brennan, 2014). The question How representative? might be answered in terms of data or sample selection procedures, such as exclusion procedures for test takers who are not part of the target population (e.g., from nonrepresentative grades) or who are nonnative English speakers or exam repeaters (Dorans et al., 2011).

Once equating data are obtained, there are additional choices about how to use the data to produce the test-linking results. The most direct choice is simply to use the sample data as originally collected to compute test-linking results, A long-standing interest in the test-equating field is the extent to which equipercentile linking results can be improved by smoothing the test data. The goal of smoothing is to reduce statistical sampling error while not substantially inducing bias. Early methods based on hand smoothing and moving averages of frequencies (Angoff, 1971; Flanagan, 1951) have been replaced with methods that are more accurate, formalized, and efficient.

One smoothing method for equipercentile linking is log-linear presmoothing (Holland & Thayer, 2000). This smoothing method is based on fitting log-linear models to the distributions of the test scores to be linked. The model fitting process produces fitted distributions that match a user-specified number of moments of the unsmoothed data. The process smooths out irregularities and provides plausible nonzero probabilities and frequencies at all possible scores. Log-linear models can be applied to the univariate distribution of the scores of a single test, to the bivariate distribution of two tests, and to the complex structure of a bivariate distribution of a test, an internal anchor and its impossible scores (i.e., structural zeros; H. Kim et al., 2017). Log-linear models can also be used to model other complexities that occur with bivariate distributions of weighted composite scores (Moses, 2014b). Because log-linear models have statistical implementations, they have associated model fit indices including chi-square statistics (e.g., likelihood ratio) and information criteria (Akaike information criterion, Bayesian information criterion, etc.) that lend themselves to statistical selection strategies (Moses, 2011; Moses & Holland, 2010). Finally, estimated variance–covariance matrices for the smoothed distributions are an output of the modeling that can be used to estimate standard errors of equating and other linking functions that explicitly reflect the smoothing results (Moses & Holland, 2008; von Davier et al., 2004a).

Another smoothing method used in equipercentile equating is known as cubic spline postsmoothing (Kolen, 1984). The "post" in postsmoothing indicates that this smoothing is implemented after an equipercentile X-to-Y equating is computed in order to directly smooth the equating function. The postsmoothing produces adjacent and connected score-level cubic functions that are numerically solved to produce the smoothest possible function that reflects the unsmoothed equipercentile function up to a user-specified smoothing constraint. The implementation, special procedures for the highest and lowest scores, and other procedures for approximating symmetry in the postsmoothed results are described by Kolen and Brennan (2014). Studies have demonstrated that cubic spline postsmoothing and alternatives are useful for reducing error in equipercentile equating functions (Cui & Kolen, 2009). Alternative strategies for selecting postsmoothing parameters have been considered (C. Liu & Kolen, 2018).

Both pre- and postsmoothing introduce bias into equating results. The logic behind both smoothing procedures is to reduce random error without substantially introducing bias. There is no definitive answer to the question, Which is better? It is clear, however, that presmoothing introduces bias before subsequent equating steps are performed, whereas postsmoothing does so when conducted after other equating steps are performed. Also, presmoothing can be automated if one adopts certain prespecified equating criteria, but doing so can be risky without examining smoothed plots before proceeding with equating. By contrast, postsmoothing involves somewhat subjective judgments about smoothing degrees.

In equating practice, data screening, presmoothing, and postsmoothing are important tools for improving the accuracy of equipercentile linking and equating functions. Reasonably large sample sizes are another important matter for ensuring accurate equating results (Dorans et al., 2011; Kolen & Brennan, 2014). When equating data are collected using small sample sizes, equipercentile functions must rely on models that make additional assumptions (e.g., pre-equating data collection designs and assumptions that work under some situations but not others; see Livingston & Kim, 2011).

Continuization

Equating functions based on scale alignment produce a set of "in between" scores that reflect plausible estimates of the difficulty differences between test forms, but that are in fact impossible to obtain. For example, if test X was to be equated to Y and X was easier than Y, then converting number-correct points that adjust for the easiness of X might result in X-to-Y scores that are lower but in between the scores that could actually be obtained (e.g., a score of 28 on X might convert to 27.6). The process of treating discrete test scores as if they are continuous is referred to as "continuizing" (von Davier et al., 2004a). The percentile rank function shown in Equation 11 is one approach to continuizing based on the assumption that the in between scores are uniformly distributed within boundaries for each score (the score ± half of the score interval). The linear scale-aligning function (Equation 16) is another approach to continuization based on standard deviation units. The continuization process in equating is described in detail by von Davier et al. (2004a), who showed that the traditional continuization approach based on percentile rank functions produces continuous distributions that can be irregular and can imprecisely reflect the original discrete distribution (i.e., overestimate variance by 1/12).

Von Davier et al. (2004a) described and developed an alternative continuization for equating based on Gaussian kernel smoothing (i.e., kernel equating; Holland & Thayer, 1989). Kernel equating is smoother, is more accurate with respect to variance, and approximates equipercentile, linear, and compromises of these functions through a user-specified kernel smoothing parameter. Several alternative continuization approaches have also been considered, including those based on log-linear models (T. Wang, 2011), exponential families (Haberman, 2011), and alternative kernel functions (Lee & von Davier, 2011). When continuization approaches work to produce values in between two adjacent, attainable scores, different approaches usually exhibit small differences from each other. They can even be shown to be iterative versions of Equation 11's percentile rank computation (Moses & Holland, 2008). One exception is when these approaches are used to equate the highest and lowest scores of a test. Equipercentile equating based on percentile rank functions will connect the highest X and Y scores to each other, whereas equatings based on kernel and linear continuizations may produce results that go further beyond the defined score ranges for the tests (Dorans et al., 2011; Kolen & Brennan, 2014; von Davier et al., 2004a). Another possible exception is scores with unequally spaced values because these are not completely suitable for Equation 11's percentile rank calculations that assume scores with ranges of integers. Flexible continuization approaches could be especially appropriate for tests with unequally spaced score ranges.

Choices Among Data Collection Designs

Each design described in the *Data Collection Designs* section and Table 11.1 presents a particular trade-off of advantages and disadvantages in test equating. These designs can be compared with respect to statistical precision, specifically the sample size of test takers needed to achieve a particular standard error of equating. For a given level of statistical precision, the randomly equivalent groups design and equipercentile equatings require more data than a single group design and linear equating (Kolen & Brennan, 2004). Designs can also be compared with respect to the complications and security concerns they introduce in test administrations (more complexity and security concerns for randomly equivalent groups designs than NEAT designs) and statistical assumptions (minimal assumptions for randomly equivalent groups, more stringent assumptions for NEAT and pre-equating designs; Kolen & Brennan, 2014). Based on the strengths of each design, a recommended ideal design is one in which large samples of test-taker data are collected in a securely implemented randomly equivalent groups design with an external anchor test that is administered after the tests (Dorans et al., 2011; Holland & Dorans, 2006). This design exploits the simplicity of the randomly equivalent groups design, but also allows for increased statistical precision through use of the anchor test as a statistical covariate. Administering the anchor externally is advantageous in that only the anchor could be affected by context or order effects, and the anchor would not need to be used if these effects were found. Pre-equating designs are usually not recommended based on their complexity and strong assumptions about IRT parameter estimates holding across multiple contexts (Davey & Lee, 2011). However, pre-equating designs could be considered for situations where a test linking must be estimated before that test is actually administered (Kolen & Brennan, 2014).

Approaches to NEAT Equating

Equating texts and studies have given extended attention to the complexities of equating using the NEAT design, considering aspects such as options and designs of the anchor test and computations based on assumptions made about the relationship of the test and anchor scores. One choice is whether the anchor will be internal or external to the tests being equated (Dorans et al., 2011; Holland & Dorans, 2006). External anchors are separately timed, and scores on them do not count toward the X and Y scores.

External anchors or sections have considerable flexibility and multiple uses, such as equating, pretesting, or the tryout of new item types. Potential drawbacks of external anchors are that they must be sufficiently disguised so that test takers do not respond differently to the anchors than they respond to the tests. In particular, test takers should not be able to determine which set of items (or test) is an external anchor, because the goal is that test takers try as hard on the external anchor as they do on X or Y. If that is not true, then scores on the anchor test will almost certainly lead to biased equating results. Recommended practices for equating with external anchors are to use data-screening procedures to identify and exclude test takers with anchor performance that is inconsistent with test performance.

Internal anchors are administered and scored within X and Y, which usually results in higher (anchor, test) correlations than external anchors, but also increases the risk of context and order effects. Recommended practices for equating with internal anchors are to administer the anchor items in similar positions on X and Y and to evaluate and possibly screen the anchor items for differential performance (i.e., differential item functioning).

The predominant recommendation is that external and internal anchors should be designed to be representative of the test, with similar average difficulties (though not necessarily equal difficulty spread; Sinharay & Holland, 2006a, 2006b). The anchor(s) should also be relatively long and reliable (Moses & Kim, 2007), where the long-standing recommendation is for at least 20 items or 20% of the items on the test (Angoff, 1971). Also, they should be administered and screened to reduce atypical test-taker performance, context, and order effects. Following these guidelines helps ensure that the anchors will be as highly correlated as possible with the tests being equated, which supports accurate NEAT equatings.

Once the NEAT data are obtained for a sample of population P that takes X and A and a sample of population Q that takes Y and A (Table 11.1), different options are available for using the anchor data to equate X and Y in T. Two major approaches involve chaining scale-aligning functions through the anchor and computing scale-aligning functions from test distributions estimated for a single hypothetical group on both X and Y.

The chained approach involves computing a scale alignment function from X to A in the P data and another scale alignment function from A to Y in the Q data and chaining them together. The assumption is that both scale alignment functions are population invariant (i.e., apply to other populations, and specifically to target population T; Holland & Dorans, 2006). Expressed in equipercentile functions (Equation 15) with percentile rank functions for A in P and Q, $H_P(a)$ and $H_Q(a)$, and linear functions (Equation 16), the chained equipercentile and chained linear scale alignment functions are

$$e_{Y,T}(x) = G_Q^{-1}(H_Q\{H_P^{-1}[F_P(x)]\}) \tag{21}$$

and

$$l_{Y,T}(x) = l_{Y,Q}\left[l_{A,P}(x)\right], \tag{22}$$

where $l_{Y,Q}[\,]$ denotes the A-to-Y linear function in the Q data.

The second approach (scale aligning) is referred to as frequency estimation or poststratification and involves projecting (see the *Distinguishing Equating From Other Forms of Linking* section) the X and Y distributions conditional on A for T. The projection is based on assumptions that the conditional $X|A$ distribution observed in population P applies to population Q and that the conditional $Y|A$ distribution observed in population Q applies to population P. The resulting X and Y distributions are estimated for a synthetic, combined group of P and Q test takers that comprise the target population $T = w_P P + w_Q Q$ ($w_P + w_Q = 1$, and w_P and w_Q range from 0 to 1),

$$Pr_T(x) = \sum\nolimits_a Pr_P(x|a)[Pr_T(a)] = \sum\nolimits_a Pr_P(x|a)[w_P Pr_P(a) + w_Q Pr_Q(a)] \tag{23}$$

and

$$Pr_T(y) = \sum\nolimits_a Pr_Q(y|a)[Pr_T(a)] = \sum\nolimits_a Pr_Q(y|a)\left[w_P Pr_P(a) + w_Q Pr_Q(a)\right]. \tag{24}$$

The X and Y distributions produced with Equations 23 and 24 are estimated for the same target population and can be used in percentile rank and scale-aligning functions like Equations 11 and 15. Different linear scale-aligning functions based on poststratification are possible, such as those that use the means and standard deviations from the distributions obtained from Equations 23 and 24 in Equation 16 (Braun & Holland, 1982). Another method attributed to Tucker (Angoff, 1971) makes assumptions that the conditional means of the tests are linear and invariant given anchor scores and that the conditional variances of the tests are constant and invariant given anchor scores, producing estimated means and variances of X and Y on T from anchor-to-test regression equations like Equation 17 (Gulliksen, 1950; Lord, 1950).

The chained and poststratification approaches to equating tests with the NEAT design have been compared in several research studies. Von Davier et al. (2004b) showed that these methods produce identical results when the (anchor, test) correlations are perfect or when the anchor distributions are identical in the P and Q samples. These results underscore the view that in ideal equating situations, such as those where (anchor, test) correlations are high and where administration groups are similar, different equating methods can produce very similar results (Dorans et al., 2011). Studies have also considered situations where the chained and poststratification methods give different results, where a choice of equating method is more consequential for reported scores. Summaries of these studies indicate that the poststratification approaches have smaller standard errors than the chained approaches and less overall equating error when group differences are small, whereas the chained approaches are less biased and have less overall equating error when group differences are not small (Dorans & Puhan, 2017; Kolen & Brennan, 2014; Kolen & Lee, 2011, 2012, 2014, 2016, 2018).

True Scores

The equating of true scores is a long-standing theoretical interest. This interest involves equating procedures based on theory and measurement models, especially classical test theory (see the *Equating Requirements* section, Equation 19; Angoff, 1971; Levine, 1955). For the NEAT design, the test forms and anchors are usually assumed to follow a classical congeneric model, and versions of Equation 19 are used to compute linear τ_X -to-τ_A and τ_A - to -τ_Y functions that are chained together for a Levine equating of τ_X and τ_Y (Hanson, 1991; Kolen & Brennan, 2014; Levine, 1955),

$$l_{\tau_Y}(\tau_X) = l_{\tau_Y}[l_{\tau_A}(\tau_X)], \tag{25}$$

where $l_{\tau_Y}[\,]$ denotes the $\tau_A - \text{to} - \tau_Y$ linear function. The relationships of the true test and anchor scores are assumed to be population invariant (Holland & Dorans, 2006).

Equating functions for true scores address theoretical interests while introducing conceptual difficulties for practice. The main difficulty is that true scores are unobserved and unavailable in practice (Lord, 1980; Lord & Wingersky, 1984) so that true score equating functions like Equation 25 typically use observed scores, $l_{\tau_Y}(x)$. Although the interchangeability claims that true score equating functions support apply to true scores, the equity discussions described in the *Equating Requirements* section show that specific types of equity are achieved in observed score applications when test forms and anchors are classically congeneric. Still, such applications are inconsistent, lack compelling reasons (Kolen & Brennan, 2014), and do not reflect the measurement error in X that is unaccounted for when equating is attempted with true score conversions (especially beyond the conditional means addressed in first-order equity). From the *Equating Requirements* section, the strongest connections in theory and practice are achieved when the test forms and anchors are constructed and administered in the same way (reflecting congeneric theories and first-order equity) and the test forms are equally and highly reliable (such that observed scores approximate true scores and $l_{\tau_Y}(x)$ functions reflect second-order equity).

Another version of Levine linear equating is based on the same assumptions for X, Y, and A and their true scores as used in Equation 25, but equates the observed X and Y scores in a hypothetical target population (defined with respect to true anchor scores) using frequency estimation and poststratification assumptions about X and Y scores given the true anchor scores (Holland & Dorans, 2006; Kolen & Brennan, 2014; Levine, 1955). Levine equating has been shown to perform well in some situations where administration group differences are large (Kolen & Brennan, 2014; Mroch et al., 2009), though these results depend on the accuracy of the Levine assumptions.

Equipercentile scale-aligning functions based on true scores have been developed by Chen et al. (2011) and T. Wang and Brennan (2008). T. Wang and Brennan's (2008) proposal is to modify the frequency estimation equipercentile procedure in Equations 23 and 24 so that these equations are based on a projection from estimated true anchor scores (i.e., $Pr_P(x|\hat{\tau}_A)$). Chen et al.'s (2011) approach generalizes chained and post-stratification equipercentile approaches developed in the kernel equating framework so that these are based on true scores.

IRT

IRT can also be used to produce equating functions of observed or true test scores or θ estimates. IRT applications require that the IRT assumptions hold and that parameter estimates for X and Y items are on a common scale. When these item parameters for the X and Y tests are estimated with a single population design such as the randomly equivalent groups data collection design (Table 11.1) using the same specification of the θ scale, the parameter estimates can be interpreted as being on the same scale. When tests are administered to nonequivalent groups as in the NEAT design, procedures are needed to transform the parameter estimates to a common scale, either through special linear transformations based on the two sets of IRT parameter estimates for the anchor items or through concurrent calibrations of the X and Y test items and anchor items (Kolen & Brennan, 2014).

For IRT observed score equating, an IRT-based recursive algorithm can be used to estimate probability distributions for X and Y given θ, $Pr(x|\theta)$, and $Pr(y|\theta)$ (Lord & Wingersky, 1984). Once obtained, these conditional distributions are averaged over a target distribution of θ; specifically,

$$Pr_T(x) = \int Pr(x|\theta) Pr_T(\theta) d\theta \tag{26}$$

and

$$Pr_T(y) = \int Pr(y|\theta) Pr_T(\theta) d\theta. \tag{27}$$

Equations 26 and 27 can be described as IRT-based presmoothed distributions (Holland & Dorans, 2006), which can be used in equipercentile equating. IRT-based true score equatings can also be produced (see the *Distinguishing Equating From Other Forms of Linking* section). Finally, if a testing program's scale scores are produced using θ as the test performance measure (see the *Scaling Methods for Primary Scale Scores* section), equating procedures may be bypassed altogether and test takers' $\hat{\theta}$s may be estimated by applying IRT ability estimation procedures to their item-level performance (see the *IRT Ability Estimates* section) and scaling these estimates to the reporting scale, $sc(\hat{\theta})$.

Accuracy Evaluations

Once estimated, test equating functions should be evaluated for accuracy. Most of the accuracy evaluations reviewed in this section are comparative, where an equating function based on one method and associated assumptions is evaluated in comparison to one or more other equating functions based on different methods and assumptions. For example, an equipercentile equating function might be of interest and evaluated in terms of how different it is from a simpler equating approach, such as a linear equating, or the use of raw scores to consider whether equating is needed at all. In subpopulation invariance evaluations, equating functions obtained from subgroup data are typically compared to a function based on the total group, usually for purposes of evaluating subgroup dependencies. These comparisons require the estimation of two or more plausible equating functions and evaluating whether their differences are "large," based on material covered in the following sections: *Differences That Matter, Standard Errors, Subpopulation Invariance Evaluations, Correlations and Prediction Error*, and/or *Evaluations and Recommendations for Drift in Equating Chains*.

Differences That Matter

Evaluations of the magnitude of equated score differences are based on whether those differences are big enough to alter test takers' reported scores. Criteria for these differences, known as "differences that matter" (Dorans & Feigenbaum, 1994), are traditionally defined based on the differences that are considered so large that they are not eliminated when reported scores are rounded, such as half of the integer of the equated scores (0.5 points for equatings based on raw summed scores) or half of the unit of the scale score interval. These values can provide benchmarks for evaluating the magnitude of equated score differences based on two equating functions, but because very similar equated scores can round to different values, the differences-that-matter criteria should not be treated dogmatically (Kolen & Brennan, 2014). Clarifications were provided by J. Liu and Dorans (2012), who proposed a procedure for evaluating whether test takers who take a test under altered conditions should be scored based on a conversion specifically produced to reflect those altered conditions. Liu and Dorans's procedure distinguished four categories of equated score differences, comprising a 2

× 2 table of rounded reported scores (are they the same or different for two possible conversions?) and the magnitude of the differences in the unrounded scores (are they trivial or nontrivial?). Based on these distinctions, Liu and Dorans recommended that a separate equating for scores based on the altered test conditions should be considered only when more than half of the test takers whose reported scores are altered (when the separate conversion is applied) also have nontrivial differences in their unrounded scores.

Standard Errors

Standard errors of equating can be useful for evaluating the extent to which potential equating functions differ by more than can be attributed to test-taker sampling errors. Standard errors of linear equating functions have a history that can be traced to Lord (1950), with updated discussions reflective of equipercentile and kernel functions, different data collection designs, distributional assumptions, and log-linear presmoothing models (Angoff, 1971; Braun & Holland, 1982; Jarjoura & Kolen, 1985; Liou & Cheng, 1995; Lord, 1982; Moses & Holland, 2008; Moses & Zhang, 2011; Ogasawara, 2001, 2003; von Davier et al., 2004a). The standard errors of equating are defined relative to the standard deviation of the distribution of equated scores produced from random samples of test takers from their respective populations,

$$SEE_{Y,T}(x) = \sigma\left[\hat{e}_{Y,T}(x)\right]. \tag{28}$$

Standard errors of equating are usually based on assumptions that the tests, their items, test-taker sample size(s), and other decisions made in the equating (e.g., smoothing) are fixed and not varied across replications of the test-taker sampling. Most of the cited sources present standard errors that are asymptotically derived, though standard errors can also be estimated through simulations and resampling approaches (Kolen & Brennan, 2014).

Standard errors for a single equating function might be used in evaluations of equated versus raw score differences, essentially questioning whether test score equating affects test scores more than can be attributed to statistical error from the test-taker sample(s). One application is estimating the test-taker sample size(s) needed to obtain a given level of equating precision (Kolen & Brennan, 2014, pp. 273–276). Another application is the situation where a testing program might offer different editions of its test in which the items, item sets, and/or test sections appear in different orders. The extent to which order effects are observed in test scores can be evaluated through equating the scores of the alternate editions to each other and evaluating how different the equated and nonequated test scores are. To the extent that the differences are "large," the equating function used to evaluate the item order effects might also be used to adjust scores for the order effects. Standard errors for an equating function provide a basis for evaluating whether equated and nonequated scores differ more than can be attributed to test-taker sampling. These ideas were used to evaluate order effects in the SAT (Dorans & Lawrence, 1990) and in Advanced Placement (Moses et al., 2007).

For questions about differences in equated scores from two plausible equating functions, each of which reflects sampling error, standard errors of equating differences (SEEDs) have been developed (Moses & Holland, 2008; Moses & Zhang, 2011; Moses et al., 2010; von Davier et al., 2004a),

$$SEED_{Y,T,1,2}(x) = \sigma\left[\hat{e}_{Y,T,1}(x) - \hat{e}_{Y,T,2}(x)\right]. \tag{29}$$

SEEDs can be used to evaluate differences between two equating functions, addressing several possible questions. Equated score differences of interest include curvilinear and linear kernel equating functions (von Davier et al., 2004a), traditional equipercentile and linear equating functions (Moses & Zhang, 2011), poststratification equating functions obtained from different target populations (von Davier et al., 2004a), counterbalanced design equating functions based on different weights of the samples taking the tests in different orders (von Davier et al., 2004a), poststratification equipercentile and chained equipercentile equating functions (von Davier et al., 2004a; Moses & Holland, 2008), and poststratification equating functions based on one or two anchor scores (Moses et al., 2010).

Subpopulation Invariance Evaluations

The subpopulation invariance requirement of equating (see the *Equating Requirements* section; Equation 20) as described by Dorans and Holland (2000) prompted several empirical evaluations of the sensitivity of linking results with respect to subpopulations (Dorans, 2004a; von Davier & Liu; 2007). Recent studies indicate renewed interest in a long-standing question in test equating (Kolen, 2004), one that can be traced to Flanagan's (1951) statements that group dependencies in equating are to be expected (p. 748) and to Angoff's (1971) statements that equating functions should be independent of the individuals used to compute them (p. 563). The inclusion of subpopulation invariance as a requirement for test equating corresponds to research and expectations for subpopulation invariance in equating. Empirical checks are important for determining whether test linkings exhibit desired equating properties and for evaluating intended equating results for tests that undergo transitions, such as in their specifications and possibly in their constructs (J. Liu & Dorans, 2013).

Consider the subpopulation invariance measure in Equation 20, for which Dorans and Holland (2000) developed two measures to quantify the lack of invariance in intended equating functions for an exhaustive set of mutually exclusive subpopulations:

$$RMSD(x) = \frac{\sqrt{\sum_{Tg} w_g \left[e_{Y,Tg}(x) - e_{Y,T}(x) \right]^2}}{\sigma_{Y,T}} \tag{30}$$

and

$$REMSD = \frac{\sqrt{\sum_{Tg} w_g \boldsymbol{E}\left\{\left[e_{Y,Tg}(x) - e_{Y,T}(x) \right]^2\right\}}}{\sigma_{Y,T}}, \tag{31}$$

where the $\mathbf{E}$ denotes an expected value and where w_g is the proportional sample size for T_g in T. The basis of the $RMSD(x)$ and $REMSD$ measures is quantifying the lack of subpopulation invariance in test equating functions of T_g from that of T, at the individual score level and also at an overall level. In their discussion of the $RMSD(x)$ and $REMSD$ measures, Kolen and Brennan (2014) suggested that these measures could mask other issues, such as the extent to which the test linking results for two subpopulations differ from each other. They proposed several additional measures that could directly address this question based on examining

$$e_{Y,T1}(x) - e_{Y,T2}(x). \tag{32}$$

Additional measures of invariance can also be computed for specific subgroups of interest (J. Liu & Dorans, 2013),

$$RESD_{Tg} = \sqrt{\sum_x Pr_{Tg}(x)\left[e_{Y,Tg}(x) - e_{Y,T}(x)\right]^2}, \tag{33}$$

where $Pr_{Tg}(x)$ is the relative frequency at $X = x$ for subgroup T_g.

Subpopulation invariance investigations were conducted using Equations 30–33 and other measures in Dorans (2004a), von Davier and Liu (2007), Kolen and Brennan (2014), J. Liu et al. (2010), and Yin et al. (2004). Standard errors for these measures were developed and studied by Moses (2008) and Rijmen et al. (2009).

One suggestion for practice is to apply subpopulation invariance concepts to conduct population invariance evaluations that address invariance for populations that are intended applications of an equating function but not necessarily represented in the actual equating study. Experience suggests that some of the most serious problems with equating and linking results involve estimating linking functions in data from one group (e.g., special studies, small subgroups of an administration used for randomly equivalent groups designs, pre-equated conversions) and applying these linking functions to other groups, testing conditions, and administration data. These experiences led to recommendations that equatings and linkings conducted in special studies with nonrepresentative scoring, administration conditions, and test-taker data should be regarded as limited in the score interpretations they suggest and subject to additional research to support their use and interpretations with the general testing population(s) (W. Lee & Brennan, 2021; Moses, 2022). This additional research can include invariance investigations for different groups as well as special data reviews that ensure that the test equating is working reasonably well for particular users and administration groups. These checks could be as simple as reviewing the scale score distribution(s) from applying an estimated equating function to the administration group and comparing it to historical distributions for that group.

Correlations and Prediction Error

Equating and other linking functions can be produced to support multiple purposes and interpretations, such as to align test scales based on score distributions and to provide predictions of test takers' scores. The scale-aligning computations used to support interchangeability across alternate forms in equating, or comparability across distinct tests in concordance studies (see the *Concordances* section), might have other intended uses as predictions of test takers' performance on those alternate forms or tests. To the extent that test correlations and reliabilities are high (see the *Distinguishing Equating From Other Forms of Linking* section), scale-aligning functions are more likely to support both aligned scales and accurate predictions. As test reliabilities and correlations decrease, the differences between scale-aligning functions and prediction functions increase, such that the predictions from scale-aligning results are increasingly biased. If the scales of independent random numbers were aligned, the results would have no predictive utility because they would not reflect the reliabilities and correlations of the "scores" (Dorans, 2004b).

When describing these issues, Dorans (1999) used a coefficient of alienation based on the prediction error of regression functions,

$$\frac{\sigma_{Y,T} - \sigma_{Y,T}\sqrt{1-\rho^{2}_{XY,T}}}{\sigma_{Y,T}} = 1 - \sqrt{1-\rho^{2}_{XY,T}}\,. \tag{34}$$

In words, Equation 34 indicates the proportion that uncertainty is reduced from predicting Y from using a linear X-to-Y regression function compared to using the mean of Y. Dorans and colleague (Dorans, 1999; Dorans & Walker, 2007) argued that concordances should only be produced when scores for tests X and Y are correlated at least .866, in which case Equation 34 indicates a 50% reduction in prediction uncertainty in standard deviation units using an X-to-Y regression. For tests correlated less than .866, scale-aligning methods result in biased predictions and should be replaced with prediction or projection methods (Dorans, 1999, 2004b). Although the correlation in Equation 34 suggests the need for a data collection design where test takers take both X and Y, Equation 34 can be used to evaluate equating results from other designs where X and Y are alternate forms assumed to be parallel. For the equating of alternate forms, the XY correlation can be estimated from the test reliabilities (Gulliksen, 1950; Lord & Novick, 1968).

These discussions were updated by Moses (2014a), who provided analogues to Equation 34 based on the prediction error from a linear scale-aligning function (Equation 16),

$$\frac{\sigma_{Y,T} - \sigma_{Y,T}\sqrt{2[1-\rho_{XY,T}]}}{\sigma_{Y,T}} = 1 - \sqrt{2[1-\rho_{XY,T}]}. \tag{35}$$

Equation 35 indicates that equatings, concordances, and other linkings produced from linear scale-aligning functions reduce prediction uncertainty by 50% when X and Y are correlated at least .875. Results were provided by Moses (2014a) to compare linear regression and scale alignment functions in terms of prediction error, showing that these functions are similar for high XY correlations but that linear regressions are increasingly favored when correlations between X and Y are low. Additional insight can be obtained by decomposing prediction error variances for regression and scale alignment functions into their proportional contributions of true score and error variances (Moses, 2014a) and by using other related measures of proportional reductions in mean-square error (Dorans, 2022).

EVALUATIONS AND RECOMMENDATIONS FOR DRIFT IN EQUATING CHAINS Established testing programs must produce several editions of a test that are developed, administered, equated, and used to score test takers. This history means that testing programs can have several previously administered and equated forms potentially available for the equating of additional test forms. Such programs would also have historical expectations about the reasonableness of scale score distributions for given administration groups. An increasing number of test equatings back to the scale is expected to cause drift due to the accumulation of random error (Kolen & Brennan, 2014). Large numbers of previous equatings also provide opportunities for monitoring historical equating results and for evaluating the consistency of current equating and scale score results.

One way that testing programs can improve test equating accuracy is to conduct their equatings to link their tests back to two or more previously equated forms (Holland & Dorans, 2006; Kolen & Brennan, 2014). From 1994 to 2016, the SAT conducted external anchor equatings to link new forms back to four old forms. Multiple links to past forms provide ways to detect aberrant equating results and scale score conversions. Multiple links also make the final equating and scale score conversions less reliant on any one previously developed equating that may be problematic.

Established testing programs can also monitor their current and historical equating results for evidence of scale score drift. Modu and Stern (1975) monitored SAT scales from 1963 to 1973, finding evidence that the verbal and mathematics sections had drifted. More recently, Haberman and Dorans (2011) delineated several contributors to scale score inconsistency, including anchors for the NEAT design, sampling errors (random and nonrandom), accumulated random error, and model misfit. Some practices were noted to exacerbate scale drift, such as continuous testing where more new forms are administered and equated with smaller groups of test takers (increasing standard errors and the accumulation of standard errors in chains of equatings). When equatings and raw-to-scale conversions are available for several administered forms, additional methods for evaluating drift have been described, such as harmonic regression and time series analyses for evaluating seasonality effects in scale score conversions (Y. Lee & Haberman, 2013) and quality control charts and time series methods to support continuous monitoring, adjustment of variations, identification of abrupt shifts, and the assessment of autocorrelation. More recent evaluations of scale stability in the SAT have also been conducted, such as studies of the extent to which drift in equating chains is affected by different degrees of postsmoothing (S. Y. Kim et al., 2020) and the drift due to the accumulation of random and nonrandom errors observed from readministering and re-equating an old form and comparing this to more recently equated forms (Guo et al., 2012). Another study evaluated the extent to which scale score drift and variability in the SAT were controlled by equating back to two or three old forms rather than one (J. Liu et al., 2014). These studies provide examples of how equating drift might be assessed and potentially reduced.

LINKING

In the *Equating* section, equating was described as the strongest type of linking, subject to the strictest requirements, used to adjust the scores of alternate forms of a single test for unintended difficulty differences for the purpose of establishing interchangeable scores. The *Linking* section describes other types of linkings for the scores of tests that are distinct and not expected to meet all of the equating requirements of equal constructs, same test specifications, equal and high reliabilities, same administration conditions, etc. Linking efforts might be undertaken to promote particular interpretations, such as appropriate comparisons or predictions from distinct tests. These goals are not as ambitious as the interchangeability goal of equating. Although the comparability established in linking might be general in intention (e.g., score comparisons across different tests and testing contexts), the results of linking distinct tests are nevertheless weaker (i.e., less precise) than equated results, with interpretations that must usually be qualified and limited in some way. For example, the score comparability achieved through scale aligning methods is considered to be "assured only for that specific group taking the tests under specific conditions (Angoff, 1971)" (Pommerich, 2016, p. 117). Predictions and regressions are also noted to be group specific (Linn, 1993; Mislevy, 1992).

This section summarizes types of linking other than equating, their uses, and their limitations. The linking types covered in earlier linking frameworks include those in the *Concordances, Vertical Scaling, Battery Scaling and Composites*, and *Predicting and Projecting* sections, all of which are summarized in Table 11.3 based on a tabled version of Holland and Dorans's (2006) linking framework. Some linking types described in previous discussions are omitted because of ambiguities and updates in terminology (*calibration* was part of earlier linking frameworks, but is avoided here because it sometimes refers to vertical scaling and other times to fitting IRT models; Holland & Dorans, 2006). Approaches described in the *More Recent Linking Types* section are also covered, including linking for exams administered in different testing modes, linking subjectively scored tests, local linking, linking adaptive tests, linking exams administered to different test takers using weak anchors, and linking state tests to NAEP (Table 11.4).

Concordances

Concordances are scale-aligning linkings for tests built from similar but not identical test specifications, where the tests have similar uses, lengths and reliabilities. One

Table 11.3 Summary of Nonequating Types of Linking

Type of Linking	What Is Dissimilar?	What Is Similar?	What Is Equal?	Likely Data Collection Design	Computation	Interpretations
Concordance		Constructs; reliability; difficulty	Population	Single group; approximated counterbalanced	Scale aligning	Comparable scales for the overall concordance group, but not necessarily to subgroups
Vertical scaling	Difficulty; populations	Constructs; reliability		NEAT with common items or scaling test, or random groups	Scale aligning	Comparable scales that depict growth based on the methodological choices for how the vertical scales were established
Battery scaling	Constructs		Population	Single group or randomly equivalent groups		Comparable scales that indicate test takers' relative performance across the test battery
Composite scales			Population	Single group taking the individual tests	The composite is a sum of the scale scores of the individual tests, which are maintained through scale-aligning equatings	Composite scales are indirectly maintained to the extent that the scales of the individual tests are maintained and the intercorrelations of the individual tests do not change
Predicting and projecting			Population	Single group; counterbalanced	Predicting and projecting	Best prediction for a prediction group

Table 11.4 Summary of Recent Linking Examples

Type of Linking	What Is Dissimilar?	Likely Data Collection Design	Computation	Interpretations
Linking across conditions of measurement	Conditions of measurement	Various	Scale aligning	Comparable scales, with caveats about limitations when testing mode differences are large
Linking subjectively scored tests	Scoring that drifts	Various	Scale aligning	Comparable scales, depending on high reliability and procedures to account for trends and rater drift across administrations

(continued)

Table 11.4 *(continued)*

Type of Linking	What Is Dissimilar?	Likely Data Collection Design	Computation	Interpretations
Local, ability-specific test linking	A form's scale alignments at different abilities	Various	Conditional scale aligning	θ-Conditional comparable scales
Linking adaptive tests	Forms and form difficulties at different abilities	Pre-equating	IRT ability estimation	Comparable scales at specific levels of adaption, influenced by the item pools, and the accuracy of the IRT models, scores and adaptions
Linking tests using weak anchors	Test-taker groups; collateral information vs. the tests	NEAT design without a suitable anchor	Scale aligning on projected distributions	Comparable scales for hypothetical groups defined by the weak anchor(s)
Linking state tests to NAEP	State tests and NAEP	Single group or counter-balanced	Scale aligning or projecting	Comparable scales or projected results are more supported within states than across them

Note. IRT = item response theory; NAEP = National Assessment of Educational Progress; NEAT = nonequivalent groups with anchor testing.

example is the concordance of the SAT and ACT reporting scales. Other concordance discussions and examples involve the ACT and ITED tests (Yin et al., 2004), and, arguably, those described in the *Linking State Tests to NAEP* section. Concordances are intended to align the tests' scales using scale-aligning functions and support additional desires to use the resulting concordance tables as surrogates for test takers' scores on the test they did not take.

Early arguments about concordances for the ACT and SAT were summarized by Pommerich (2007), including concordance table proposals for avoiding excessive testing in schools and arguments from Angoff (1962) and Lindquist (1964, February) against concordance tables because of problems in these proposals and possible misuses of their results. Since these early arguments, several ACT/SAT concordances have been produced (College Board/ACT, 2018; Dorans, 1999; Dorans et al., 1997; J. Liu et al., 2010; Marco & Abdel-Fattah, 1991). These ACT/SAT concordance studies are based on obtaining data from test takers taking the ACT and the SAT at least once, in either order, selecting the ACT and SAT scores to use for the test takers who take either test more than once and screening the test-taker data to avoid large time differences between ACT and SAT testings. Linking results are produced from the screened data as equipercentile conversions for some combinations of ACT and SAT tests (e.g., the ACT Summed score and the SAT Total scores in 2018, the ACT Summed score and the SAT Verbal + Math score in 2010). Decisions about which ACT and SAT tests to concord are usually based, in part, on the size of their correlations (see the *Correlations and Prediction Error* section; Dorans, 1999). Content similarities and political issues often inform these decisions as well. The usual data collection design is a single group design that ignores testing order. In some ACT/SAT concordance studies, testing order is accounted for by weighting data for each order similar to a counterbalanced design, but for nonequivalent groups (College Board/ACT, 2018; Marco & Abdel-Fattah, 1991). Other statistical weighting procedures have also been used to improve the representativeness of the concordance data and resulting concordances for the test-taker populations of each test, who do not necessarily take both tests (College Board/ACT, 2018).

Because of differences in the specifications of the tests in concordance studies, concordance results are not expected to provide interchangeable scores. In fact, concordance tables are expected to be group dependent. For example, the concordance of the ACT Summed score and the SAT Verbal + Math score differs for females and males and across different race/ethnicity groups. These differences can be attributed in large part to the fact that math content contributed 25% to the ACT Summed score and 50% to the SAT Verbal + Math score (Dorans, 2020). In addition, concordances produced from very specialized conditions of measurement, scoring, or with restricted samples or in special studies may not generalize to entire testing populations and may warrant additional studies to update results (Dorans & Moses, 2023; W. Lee & Brennan, 2021; Moses, 2022).

Vertical Scaling

Vertical scaling is used to establish "developmental" scales for reporting performance on versions of a test that are appropriate for specific ages or grade levels represented in schooling, and usually K–12 testing (Kolen, 2006; Kolen & Brennan, 2014). In Holland and Dorans's (2006) linking framework, vertical scaling involves the linking of tests that measure similar constructs at similar reliabilities, but that differ in difficulty and in the test-taker populations. Thurstone (1925, 1938) proposed vertical scaling methods for item difficulties and then modified them for summed scores (later described as a method of absolute scaling for age and grade-based scales; Flanagan, 1951; Gulliksen, 1950). Rasch IRT models were considered for vertical scales in the 1970s and 1980s (Briggs & Weeks, 2009). More recent vertical scales include CTB/McGraw–Hill's Terra Nova (CTB/McGraw–Hill, 2001), the Iowa Test of Basic Skills (Hoover et al., 2003; Kolen 2006; Petersen et al., 1989), the ACT scales (Brennan, 1989), tests of English acquisition and English as a second language (ETS, 2005; J. Wang & Smith, 2003), and the 2016 redesigned SAT Suite (Y. K. Kim et al., 2016). Typically, a separate vertical scale is established for each test in an overall testing battery, serving as potential inputs to battery scaling (see the *Scales for Test Batteries and Composites* and *Battery Scaling and Composites* sections). The resulting developmental scales and scores are used to provide a means by which students' growth is measured, observed, and used to plan instruction and instructional support across schooling. Because the tests being linked in the vertical scale differ in difficulty, length (usually), timing, content, and other aspects, vertical scaling results do not produce interchangeable scores across levels (Kolen & Brennan, 2014).

Several issues affect the production and results of vertical scales (Kolen, 2006), including definitions of growth, grade-based testing content, data collection designs, linking methodologies, and implementation choices in an IRT model, scoring, and estimation.

How Growth Is Defined

Vertical scalings can reflect different definitions of growth, such as *grade to grade*, where growth is defined over the test content appropriate to particular grades (usually two adjacent grades), and *domain*, referring to growth over all content in the domain (Kolen & Brennan, 2014).

Grade-Based Testing Content

For subject matter areas closely tied to a school curriculum, students tend to exhibit different amounts of growth depending on the content areas on which students are tested. For example, if division by whole numbers is taught in Grades 3 and 4, then growth in the third and fourth graders is expected to be greater in this area than growth in fifth and sixth graders (Kolen & Brennan, 2014). Different vertical scaling results might result from conducting the scaling on either grade (Holland, 2007, p. 18). Growth rates would be different across vertically scaled tests measuring domains that are differentially associated with school curriculum. Vertical scaling results can also be affected by how well the difficulty levels are represented in tests for earlier and later grades (Kolen, 2006, p. 178).

Data Collection Designs

Three designs are usually contrasted for vertical scaling (Kolen & Brennan, 2014). For vertical scales established with the NEAT design, adjacent grade-based tests are taken by students from those corresponding grades, and the blocks of items representing overlapping content for those two grades serve as the anchor test. The randomly equivalent groups design can also be used, where test takers in each grade are randomly assigned to take either a grade-specific test or a test designed for an adjacent grade. The NEAT and equivalent groups designs both involve the chaining of test linkings for tests administered to adjacent grades, and both designs reflect a grade-to-grade definition of growth.

A third data collection design is the scaling test design, which is similar to the NEAT design where an external "scaling" test that is built to reflect content from the entire domain and represent all of the grades of interest is used as the anchor. Typically, the scaling test is administered to students along with the grade-specific tests to be scaled. The scaling test design reflects a domain definition of growth. The scaling test design has been used with the Iowa Test of Basic Skills, ACT, and SAT testing programs.

Linking Methodologies

Vertical scales can make use of methods based on several measures of test performance (Kolen, 2006). For summed scores, Hieronymous scaling involves the linking of summed score medians across tests. Thurstonian scaling is based on linking grade-based summed score distributions after they are transformed into normal distributions. The untransformed raw scores could also be vertically scaled, such as through chained equipercentile procedures (Y. K. Kim et al., 2016). IRT scoring methods can also be used to establish vertical scales.

IRT Implementations

The use of IRT for vertical scaling involves choices of an IRT model, estimation choices, and the IRT-based scores to use (Briggs & Domingue, 2013; Briggs & Weeks, 2009; Kolen, 2006). IRT models such as the Rasch, 2PL, or 3PL models could be used. Test items' IRT parameter estimates could be obtained through calibrations conducted separately for each grade-specific test, items, and test-taker data or concurrently for the tests, items, and test takers from all grades. Vertical scales could be established from different proficiency estimators, including MLE (Equation 4), EAP (Equation 5), or TCC approaches (Equation 6).

The test score linkings resulting from vertical scales have been characterized as a "folding ruler" and as a "ruler that bends" (Yen, 2007, pp. 274–275). What causes vertical scales to fold and bend? As noted previously, the grade-specific tests can show more growth for content that is specifically taught in the curriculum for those grades (perhaps suggestive of grade-to-grade growth). Reviews of vertical scales from the 1980s describe a shrinkage phenomenon where the scores from several types of vertical scales tended to have standard deviations that shrunk from earlier to recent tests and grades (Briggs, 2013; Kolen, 2006; Yen, 2007). Shrinkage was most often shown in vertical scales established with IRT, potentially affecting the ordering of schools based on their gain scores (Briggs & Domingue, 2013). Shrinkage in IRT scales results in a depiction of growth that differs from the growth depicted in vertical scales established with non-IRT approaches, where shrinkage suggests more rapid growth for lower achieving students who catch up to higher achieving students (Briggs, 2013; Hoover, 1984). Although the source of the scale shrinkage in vertical scales from the 1980s was never definitively established, suggested explanations point to early IRT estimation methods (i.e., joint maximum likelihood estimation), multidimensionality within and across the tests, and a failure to establish interval scales. Vertical scales since the 1980s have generally not exhibited shrinkage (Yen, 2007), coinciding with updates in IRT procedures.

Additional influences on vertical scales have been elaborated (Briggs & Weeks, 2009), including the chosen IRT model (scales based on 3PL models result in greater scale score variability than those based on Rasch models), IRT estimation (MLE results in greater score variability than EAP), and calibration (concurrent calibration decreases scale variability vs. separate calibration). IRT limitations can also affect vertical scales, such as strong and potentially unrealistic assumptions that item parameter estimates fit item response data and are invariant across ages and grades, as well as assumptions that items and test-taker abilities can be modeled as unidimensional. Multidimensional approaches that account for changes in constructs measured across grades can show limitations with unidimensional approaches (Weeks, 2018). For these and other issues described in the cited sources, vertically scaled scores across grade- and age-specific tests can produce inconsistent estimates of test takers' growth, estimates that reflect choices in methodology, test development, and administrations. Altogether, these results indicate that scores from different tests in the vertical scale are not interchangeable.

Battery Scaling and Composites

Battery scaling can be described both as an approach to establishing scales (see the *Scales for Test Batteries and Composites* section) and as an indirect way to link the scores of a battery of tests designed to measure different constructs that are administered to a common population of test takers (Holland & Dorans, 2006). The purpose of establishing the scales in a similar way for all the tests in a battery is so the resulting scales can facilitate interpretations about relative performance and strengths and weaknesses across the battery. Examples include an early proposed battery scale for different scoring methods for handwriting (Kelley, 1914) and for establishing scales for recent versions of the SAT and ACT batteries (Brennan, 1989; Dorans, 2002; Y. K. Kim et al., 2016). It is likely and expected that the linkings of test scores produced in battery scaling do not produce interchangeable test scores. One way that this lack of interchangeability can be observed is in subpopulation dependencies, such as for subgroups of test takers who do relatively better (or worse) on a mathematical measure and also do worse (or better) on a verbal measure.

Composite scores are usually derived from individual tests in a battery, making the comparability of composite score scales indirect and difficult to maintain. To illustrate some challenges with composite score scales, consider a situation where the scales of two tests are established and maintained through equating, and let these two tests for which alternate forms would be developed be represented as Y_1 and Y_2. A composite score is also of interest, defined here as the sum of the two scale scores, $sc(Y) = sc(Y_1) + sc(Y_2)$. For this situation, the distribution of $sc(Y)$ is a function of the scale score distributions of $sc(Y_1)$ and $sc(Y_2)$ and also the joint distribution of $[\,sc(Y_1), sc(Y_2)\,]$. Specifically, the mean of $sc(Y)$ can be obtained as the sum of the means of both tests' scale scores, the variance of $sc(Y)$ is obtained as the sum of the variances of both tests' scale scores plus two times their covariance, and the skewness and higher moments reflect higher moments in $sc(Y_1)$ and $sc(Y_2)$ and in the joint distribution of $[\,sc(Y_1), sc(Y_2)\,]$. These relationships indicate that except for its mean, the distribution of the composite scale scores $sc(Y)$ is a function of the $sc(Y_1)$ and $sc(Y_2)$ scale scores that would be maintained through equatings of the Y_1 test forms to each other and the Y_2 test forms to each other, as well as the covariance and higher moments of the joint distributions of $[\,sc(Y_1), sc\ (Y_2)\,]$that would not be maintained through the Y_1 equatings and the Y_2 equatings. Examples of composites produced from multiple tests include those reported by ACT, SAT, and other programs. Kolen and Brennan (2014) summarized how a change in composite score standard deviation coincided with the change in intercorrelations of separate battery tests' scale scores and recommended specific checks for composite score comparability: "When composites are created for tests in a battery, it is important to check whether the composites are also comparable" (p. 425). Analogous results have been shown in linking situations involving latent composites (Dorans et al., 2014; Weeks, 2018). An implication is that composite score scales are best maintained directly rather than linked through indirect reliance on equated battery tests.

Predicting and Projecting

Predicting and projecting are the oldest and earliest examples of test linking (Holland & Dorans, 2006). Although these linkings are not symmetric like scale-aligning functions, various applications have been described in the psychometric literature. Predictions and projections have been recommended because of their increased prediction accuracy versus scale-aligning functions, such as when there are low correlations in the scale scores for which a concordance is desired (Dorans, 1999; 2004b), or in the IRT thetas of distinct measures intended to be linked (Holland & Hoskens, 2003; Schalet et al., 2021). Other uses are for establishing auxiliary scales, such as reports of normative growth using conditional norms (i.e., relative growth rather than the absolute growth intended to be shown in vertical scales). Conditional growth norms might be used to forecast how younger test takers obtaining specific scores on an earlier test are expected to perform on a later test (Betebenner, 2009; Castellano & Ho, 2013). For example, PSAT-to-SAT prediction tables are descriptions of expected growth at the student and school levels (Y. Kim et al., 2018a, 2018b). As described in Holland and Dorans's (2006) linking framework (also Linn, 1993; Mislevy, 1992), predictions and projections are the least restrictive type of linking in terms of requirements. They do, however, require a sample of test takers who took both tests. The usefulness and accuracy of prediction and projection tables are higher when the tables are applied to test takers who are most similar to the test takers used to produce the tables (Holland & Dorans, 2006). In other situations, the group dependences noted in prediction and projection tables can make them unstable, inaccurate, and "precarious" (Mislevy, 1992, p. 63), such as in applications to nonrepresentative groups or to groups over time (Thissen, 2007).

More Recent Linking Types

Since Holland and Dorans (2006) presented their test linking framework, different types of linking have been proposed and developed. Some prominent examples are summarized in this section and in Table 11.4. A focus of this discussion is the extent to which these more recently proposed linking types differ from those in the *Equating Requirements* section. Several of these proposals are based on desires to offer tests in more specific ways to test takers by offering tests in an increased number of administration modes (see the *Linking Tests Across Conditions of Measurement* section), supporting tests with subjective scoring procedures (see the *Linking Subjectively Scored Tests* section), or developing the tests or the intended linking function tailored to specific test takers (see the *Local, Ability-Specific Test Linking*, and *Linking Adaptive Tests* sections). Additional proposals expand on methods for linking across nonequivalent groups (see the *Linking Tests Using Weak Anchors* section) and the *Linking State Tests to NAEP section*.

Linking Tests Across Conditions of Measurement

Testing programs might allow test takers to take their tests in different administration modes, such as translated into different languages, with modifications and accommodations for special populations, and in multiple delivery systems that include paper–pencil tests and computerized tests delivered online. When offering alternative editions of their tests that are administered under different conditions, the testing program can be faced with two choices. One choice is to ignore administration effects and treat the scores from a test given in different administrations in the same way. This choice could raise fairness concerns and result in an unknown degree of comparability in scores across measurement conditions (Pommerich, 2016). The other choice is to conduct linking studies to estimate and apply score adjustments that account for administration effects on scores (i.e., mode comparability studies). This choice may be possible for evaluating score effects from computerized versus paper testing modes, but is less feasible for tests administered with and without accommodations or in different languages where standard data collection designs may be less available (Thissen, 2016).

Mode comparability studies and linkings produced in these studies do not meet the equating requirement for equal conditions of measurement (see the *Equating Requirements* section) and are not expected to produce interchangeable scores. However, these studies might improve or at least inform the fairness of score reporting. Some reviews of paper–pencil versus computerized mode comparability studies suggest that scores from these modes are comparable more often than not (Pommerich, 2016), whereas other studies suggest that computerized testing can favor some subgroups over others (Kolen & Brennan, 2014). The mixed findings could be due to challenges in conducting mode comparability studies with traditional data collection designs (see the *Data Collection Designs* section) and the extent to which the administration modes differ.

Traditional data collection designs can be challenging to implement with mode comparability studies (Kolen, 2007). Mode comparability studies with randomly equivalent groups designs require that test takers be randomly assigned to an administration mode (paper–pencil, computerized, etc.), an assignment that can differ from the typical administrations of tests and modes by a testing program. A NEAT design implementation would require that test takers take an anchor test given in a single administration mode that may differ from the mode in which they take their actual test. Single group and counterbalanced designs can produce scores with order effects (Eignor, 2007). Traditional designs have been implemented in special studies for large-scale assessments like NAEP, the Programme for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS; Jia & Xi, 2021) and have found that digitally administered items in mathematics and reading (NAEP) and math and science (TIMSS) were more difficult for fourth and eighth graders. Other studies have assessed mode effects in paper–pencil and online administrations of credentialing, licensure, and placement exams, using a variety of methods to account for test-taker nonequivalence and finding mixed results (Jones et al., 2022; S. Kim & Walker, 2021; Moses et al., 2021; Puhan & Kim, 2022).

A major question is how different the tests are when administered in each mode, such as in terms of their display of items and other test material, the test administration(s) (fixed, adaptive, etc.), and the extent of proctoring for the computerized administration (proctored at test centers vs. unproctored or online administrations). Discussions of these issues suggest that mode comparability linkings are most dissimilar with equatings that produce interchangeable scores across testing modes when the computerized test is adaptive and unlikely to exhibit equity with the paper–pencil scores (Eignor, 2007). In summarizing the transitions many admissions and placement programs made to online testing in efforts to continue administering tests in the COVID-19 pandemic, Camara (2020) noted that online administrations with especially large differences from paper–pencil tests risked the integrity, accuracy, reliability, validity, and fairness of the resulting scores and would "likely mean score trends and comparability cannot be maintained" (p. 13).

Altogether, the challenges with data collections and with potentially large differences in administration modes make mode comparability studies challenging in their implementation and their interpretation. Thissen (2016) provided a general discussion of testing and linking across different translations, grade levels, accommodations, and paper–pencil versus computerized modes, concluding that empirical investigations of fairness are needed for tests and linkings across conditions of measurement:

> Continued vigilance is required. There is no way to guarantee that a use of a test is fair. All that can be done is to catalog carefully the ways in which comparisons could be unfair, and then check, with either statistics or reasoned judgment, whether unfairness exists. (Thissen, 2016, pp. 212–213)

Linking Subjectively Scored Tests

Tests that are composed of CR items present challenges for linking, such as tests that usually contain a small number of items and data collection designs that do not work the same way as for tests containing MC questions (Kolen & Brennan, 2014). The use of human raters to score tests containing a small number of CR questions can reduce reliability. Automated essay scoring approaches may avoid reliability reductions from human rater scoring, but they can also introduce their own challenges for score interpretability and fairness (see the *Ability Estimates From More Complex Models* section). The previously described issues about unreliable scores having less utility and producing linkings with less equity indicate that linkings for tests with CR items are not likely to produce interchangeable scores. Tests with small numbers of questions can also result in inadequate coverage of the content domain, such that alternate forms may measure different constructs. Low reliability and inadequate content coverage can make equating and score interchangeability less likely for tests with CR questions.

When tests containing CR questions must be linked using a NEAT design, linking accuracy depends on a number of factors and procedures. A crucial issue with NEAT designs for CR tests is what to use for the anchor test. When the anchor contains CR questions, the scores from these anchors can reflect a particular type of bias due to leniency/stringency differences in human raters from the two administration groups (Tate, 1999, 2000; S. Kim et al., 2010a, 2010b). Trend scoring procedures are recommended for these biases, where the responses to a set of CR questions from a single group of test-takers are scored by raters in both administrations. A scale-aligning transformation for the sets of scores from these administrations is produced to account for the unintended differences in rater leniency/stringency across administrations and to produce scores that can be used more effectively as an anchor for linking the tests. The use of trend scored CR questions as an intact anchor, or as part of an anchor with MC question scores for mixed format tests, improves accuracy when linking mixed format tests with the NEAT design (S. Kim et al., 2010a, 2010b). Another choice available for mixed format tests is to use anchor tests composed only of MC items. This approach is most effective when the administration groups are similar to each other on the MC and CR sections, when the correlations of the MC and CR scores are high and similar across forms, and when the number of score points attributable to MC scores is large (Kolen & Lee, 2011, 2012, 2014, 2016, 2018).

Local, Ability-Specific Test Linking

If an equating satisfies equity, then it is a matter of indifference whether test takers at every given ability level take form X or Y (see the *Equating Requirements section*; Lord, 1980). Consider Lord's original definition of equity, defined over every ability level (θ value from an IRT model), and emphasize "each" ability level (van der Linden, 2011, p. 209; van der Linden, 2013, p. 262). A set of X-to-Y linking functions could be defined to satisfy this interpretation of equity at specific values of θ:

$$e_{Y,\theta}(x) = G_{\theta}^{-1}[F_{\theta}(x)] \text{ such that } F_{\theta}\left[e_{Y,\theta}(x)\right] = G_{\theta}[y] \quad \text{for each } \theta. \tag{36}$$

The goal of Equation 36 is to estimate each set of θ-conditional X and Y distributions (Lord & Wingersky, 1984) and use them to produce θ-conditional X-to-Y equipercentile functions. This proposal contrasts with the usual recommended practice to average the conditional distributions of X and of Y given θ (Equations 26 and 27) and conduct an equating for $e_{Y,T}(x)$ on the averaged distributions (Kolen & Brennan, 2014). If test takers' estimated θs were available from an IRT model, a set of equating functions at each estimated θ value could be used to evaluate lack of equity in the equated scores, $e_{Y,\theta}(x) - e_{Y,T}(x)$, rather than in distributions (Equation 18), analogous to evaluations of subpopulation invariance (Equation 20; Dorans & Holland, 2000).

One proposal is to use Equation 36 to produce and report θ-conditional scores to test takers, to satisfy equity and address equating bias attributable to θ-conditional measurement error (van der Linden, 2011, 2013). This proposal has been criticized for theoretical reasons and practical issues (Dorans, 2013; Holland, 2013). Some questions are discussed next to elaborate both perspectives, the proposal for θ-conditional procedures and counterarguments.

- Does the target population for equating, T, contain test takers at all θ values or is it θ-conditional? One perspective is that there are multiple θ-conditional target populations (van der Linden, 2011, p. 209). Another perspective defines T in terms of all test takers taking a test in a data collection design from Table 11.1 (i.e., test takers at all θ values). From the perspective that T contains test takers at all θ values, θ-conditional linking functions that differ from $e_{Y,T}(x)$ are an indication that the subpopulation invariance requirement is violated (Requirement e, Dorans, 2013). A θ-conditional interpretation would consider these differences to reflect different populations.
- Should administration conditions like scoring rules be applied in the same way to test takers at all θs or should different scoring procedures be applied to test takers at different θs? Unique scoring rules might be applied to test takers at specific θ values or consistently to test takers at all θ values. Note, however, that θ-conditional linking functions would imply that the equal measurement conditions requirement of equating is not satisfied (i.e., test takers with different θ values who take test X are scored differently from each other and likely from Y). Psychometricians have usually argued for the use of uniform scoring for all test takers taking a given test form (Livingston, 2004; Petersen, 2007).
- What are the implications of θ-conditional linking functions for score comparability? From one perspective, θ-conditional equipercentile functions correct for θ-conditional measurement errors (van der Linden, 2011, 2013). Another perspective is that θ-conditional equipercentile functions result in differential treatment of test takers (Dorans, 2013; Holland, 2013) and θ-conditional score interpretations that would be contradictory to interchangeability across different θ values (i.e., score comparisons across θ values are not supported).
- Is equating an adjustment for the estimated difficulties of X and Y in T or an adjustment for θ-conditional measurement error? One perspective is that the equating of observed scores involves adjusting for θ-conditional measurement errors (van der Linden, 2011, pp. 213, 223). The perspective of most other works cited in the *Equating* section is that equating involves adjusting test form scores for unintended differences in overall difficulty (see the *Equating Requirements section*), where the test forms involved are equally and highly reliable.

Rather than use Equation 36 to report θ-conditional linking functions to test takers, the recommendation in this chapter is to re-emphasize high reliability as a requirement for equating. This means that a single X-to-Y conversion applied to all test takers of X

would reflect less measurement error, would be a closer approximation of the equity requirement, and would have small differences from θ-conditional X-to-Y functions. The requirement for high and equal reliability in test forms being equated means Lord's (1980) theorem that only perfectly reliable true scores can be equated is more closely approximated. This perspective makes use of Equation 36 not as a score reporting strategy, but as one possible check for equating adequacy, namely, a check on the extent to which θ-conditional measurement error and (in)equity affect equating accuracy.

Linking Adaptive Tests

Adaptive testing is based on the premise that a test can be more precise when it is constructed so that its difficulty is matched ("tailored," Lord, 1980, p. 150) to test-taker ability. This matching of test-taker ability and test difficulty implies increased efficiency, in that a matched test can be shorter but measure at a precision that is similar to or higher than that of a longer test that may be too easy or too difficult for an unmatched test-taker group. Adaptive tests are usually IRT based in their scoring, linking, and scale scores, relying on a large pool of available test items with IRT statistics obtained from pretest and pre-equating data collections and using computer-based algorithms to administer the tests (van der Linden & Glas, 2010).

When the test is administered, a computer program implements sophisticated algorithms to obtain preliminary IRT-based ability estimates as a test taker takes test items. Then the computer program selects and administers subsequent test items with difficulties that match test takers' ability estimates. The computer might stop the test when the test taker's ability estimate reaches a predetermined level of precision. Additional procedures are needed to ensure that the item pool is adequately maintained, that items are not being overly selected or exposed so as to raise security concerns, that content specifications are met, and that items' IRT parameter estimates are not drifting during repeated use. Adaptive testing could be implemented at the item level (i.e., every subsequent item a test taker takes is based on his/her ability estimated from previously taken items). Adaptive testing could also be implemented at a small number of stages where test takers are routed to collections of items of differing difficulty (i.e., multistage tests, MSTs). Several examples of item-adaptive tests and MSTs are described by van der Linden and Glas (2010).

Computer-adaptive tests present several challenges for score linking. The adaptive tests are highly dependent on the item pools, such that systematic changes in the item pool through item reduction or the use of different exposure controls can result in test scores that are not interchangeable because of increased measurement error, violations of second-order equity, and changes in the distribution of reported scores (T. Wang & Kolen, 2001). Even when alternate item pools are built to be very similar, they might produce scores that are different enough to warrant additional linking beyond what is provided in the adaptive test (Segall, 1997). Adaptive tests rely on assumptions that scoring parameters are correct and item order and context effects are either minimal or can be controlled. More thoughtfulness has been encouraged about these assumptions (Harris, 2023), and special research has been prompted to assess pretested items administered in varied positions (Davey & Lee, 2011). Importantly, the design of an adaptive MST will reflect one of several possible trade-offs of adaption and the extent to which content specifications are met:

> More modules per stage may make a test more adaptable to a wider range of examinee proficiency levels, but then, more easy items and hard items are needed to build the MST modules. The items must also be selected to simultaneously meet all requisite content specifications for any test route taken by the examinee. This is often very challenging when the modules at a given stage must be matched on content and also span a fairly wide range of item difficulty. Similarly, fewer items per stage likewise encourages the use of more adaptation, but can result in routing decisions being made on smaller and smaller slices of the content domain. (Zenisky et al., 2010, p. 356)

Adaptive tests have some resemblances to vertical scales (see the *Vertical Scaling* section), in that they involve tests developed to differ in difficulty and possibly in content being administered to test takers of specific estimated ability. Both linking types can fail to meet the equating requirement for content and difficulty (Requirement a, see the *Equating Requirements* section) and both raise comparability challenges. Similar to how different vertical scaling results could be produced as students of different ages and/or grades take the tests being vertically scaled, different scores from computer-adaptive tests could be produced if test takers of higher and lower ability take one of several unique and adaptive tests differing in their difficulty and possibly content. These issues might be a matter of degree where, for example, simulations of MSTs that administer modules with relatively small differences in difficulty and content may not show large scoring errors in misrouted test takers (S. Kim & Moses, 2014). However, as the quote from Zenisky et al. (2010) suggests, it is possible that the adaption in adaptive testing could be more extreme, as is the case with item-level routing, such that tests targeted at specific ability levels differ from those that target other ability levels in their difficulty and content coverage. Greater degrees of adaption and larger differences in the adaptive tests administered to test takers can reduce the comparability of scores across test takers of different ability.

When unidimensional IRT models are employed, it is virtually impossible for test-taker scores to be successfully equated for adaptive "forms" because, by design, the adaptive routes differ in difficulty and indirectly modeled content, at least in part, as well as the test-taker groups that take the different routes. In effect, different ability groups take different test forms that are intentionally designed to be nonequivalent, which rules out equating as an attainable goal. (Similar statements apply to linking scores on paper–pencil and MST forms.) Just about all educational tests are multidimensional by design and by content specification. When a unidimensional IRT model is used and adaptions reflect ability, difficulty, and incompletely modeled multidimensional content, statements about scores for test takers who take different test forms and routes being on the same scale are difficult to defend. Linkings of some kind may be possible, but the characteristics of any linked scores can be ascertained only through empirical studies that consider the accuracies of the IRT parameters and scores, the item pool, the adaptive settings, test forms for higher and lower ability test takers, etc. The strong claims supportable by an actual equating are not expected in such linkings.

Linking Tests Using Weak Anchors

Consider the situation where a linking is desired for tests X and Y, but the test data are not collected in one of the recommended data collection designs (Table 11.1). If these test data were obtained in a manner similar to a NEAT design in which X and Y were taken by nonequivalent groups, but with no anchor score data, what linking approaches might be available? One example of this situation occurs when the data from a paper–pencil test were intended to be collected in a randomly equivalent groups design, but test book spiraling procedures failed. Another situation occurs when the administration of a common anchor to both groups is not feasible because of security concerns. Finally, testing programs might intentionally introduce this type of design by offering their test in different administration modes and giving test takers the option to self-select into different test administration modes.

Two issues that determine the quality of linking are the similarity/dissimilarity of the administration groups and the representativeness of the available anchor(s), (A(s)) for X and Y as indicated in the anchor, test correlations. An ideal situation occurs when the administration groups are samples from the same population. For this case, Holland and Dorans (2006) described the role of the anchor as one of reducing random variability and improving statistical precision, even when the anchor does not represent the tests:

> When $P = Q$, the NEAT design is called an EG [randomly equivalent groups] design with anchor test. The two samples are drawn from a common population and the role of the anchor test changes. The anchor test becomes a covariate as in a randomized experimental design. It is used to gain precision in the estimation of the relevant parameters, rather than to adjust for group differences. For this special case, it is not necessary for A to measure the same construct that X and Y do, or even to be a test score. All that matters is for A to be correlated with both X and Y. When this is the case A is useful as a precision-increasing covariate. (p. 199)

For less than ideal situations where the populations of the administration groups cannot be considered equivalent, the anchor must reduce bias and variability (Holland & Dorans, 2006). The potential for greater linking error when administration groups are not equivalent means that it is especially important that the anchor be representative of, and correlated with, test scores. Although the use of inadequate anchors for a linking test administered to nonequivalent groups might improve linking accuracy, this use may not sufficiently account for test-specific group differences or produce equatings that support interchangeable scores.

Several linking methods have been considered for situations where the administration groups differ and where multiple anchors and/or background variables might be available that could be incorporated into a test linking. Suggestions for using multiple scores were made by Angoff (1971) in a generalization of regression methods (Equation 17). Livingston et al. (1990) suggested propensity score matching as a way to incorporate multiple scores. Dorans and Wright (1993) showed that linking results obtained from matching administration groups based on a selection variable could be more accurate than methods based on matching groups with traditional anchors. Liou et al. (2001) described how demographic background variables might be used in test linking by making assumptions about the missing data and using corresponding imputation methods. Moses et al. (2010) showed how frequency estimation based on projected distributions (see the *Distinguishing Equating From Other Forms of Linking section*), categorized propensity scores, and missing data imputation could all be used to produce similar linking results when these methods were used with two anchor scores. Studies of categorized propensity scores based on background variables (Livingston, 2014; Wallin & Wiberg, 2019) and linear scale alignments with linear regressions of multiple background variables (Bränberg & Wiberg, 2011) showed that these approaches can increase precision and reduce variability in test linking. Statistical weighting procedures based on log-linear models have also been described for test linking based on background variables (Haberman, 2015) and on test takers' previous test scores (Y. Lee et al., 2019).

Most of these methods produce scale-aligning results that are similar to those obtained from projecting test score distributions for hypothetical administration groups defined by one or more anchors and/or background variables. It is possible that the linking results based on using an anchor and/or background variable have improved accuracy compared to an inappropriate randomly equivalent groups approach. However, to the extent that the anchors and/or background variables do not represent (and correlate highly with) the tests being linked, and also do not account for administration group differences, the linking results will likely have inadequacies, such as insufficiently controlled group differences that preclude interchangeable scores.

Linking State Tests to NAEP

The National Research Council published *Uncommon Measures* (Feuer et al., 1999) to address a debate in the late 1990s between those who favored voluntary national tests as a means of assessing the educational progress of students across the nation and those who believed that statistical linkages among existing tests could be used to achieve that purpose. The *Uncommon Measures* report examined the feasibility of linking the results of commercial and state tests, such as by linking these tests to each other and to the NAEP scales to support comparisons of students' achievement with national and international benchmarks and with students in other places. Feuer et al.

demonstrated that it was not possible to link state assessments for several reasons, including differences in the NAEP and state tests with respect to their content, format, margins of error, intended and actual uses of the tests, and the consequences attached to the test results.

Attempts to link state tests to the NAEP scale have continued since Feuer et al.'s (1999) report (Dorans, 2020), prompting discussions and questions about the interpretational value of the results. Thissen (2007, 2016) discussed an approach based on projection methods (Williams et al., 1998), and listed concerns about lack of population invariance, indications that linking results were not stable over time, study cost, and the unknown levels and differences in motivation for students taking a standardized test versus the NAEP test, which is more of a survey that lacks direct personal consequences for test takers. Other attempts involve linking state standards to NAEP scales using equipercentile methods (Braun & Qian, 2007; McLaughlin, 2000; McLaughlin & Bandeira de Mello, 2002). These attempts prompted concerns about whether the inferences and interpretations were too strong, about potential instabilities of the results over time (Koretz, 2007), and about comparisons that may not be defendable for the states and state tests that differ from NAEP with respect to content frameworks, implementation, and stakes (Ho & Haertel, 2007).

In another attempt to use the NAEP scale to connect disparate state assessment results, Reardon et al. (2021) reported linear scale-aligning linkings of school district means on state tests to the NAEP scale, based on statistically inferring the district means on the state tests from published passing rate distributions. Commentaries on Reardon et al.'s article by Bolt (2021), Davison (2021), Moses and Dorans (2021), and von Davier (2021) re-emphasized long-standing cautions about the large variation in blueprints used by different state tests and other differences with NAEP in terms of content, administration conditions, stakes, and test-taker motivation. Moses and Dorans (2021) provided empirical demonstrations that state-based linkings of district means from one test to another are not invariant across states, even when correlations in the district means of these scores are very high (.98 or higher). These discussions suggest that cross-state invariance evaluations are needed to support cross-state comparisons of districts.

DISCUSSION

The reporting scales for a large-scale testing program are the focus of five chapters of the previous editions of *Educational Measurement* (Angoff, 1971; Flanagan, 1951; Holland & Dorans, 2006; Kolen, 2006; Petersen et al., 1989). The current chapter updated these discussions. The *Scaling* section covered approaches to setting the scales for the test(s) of a large-scale testing program. The *Equating* section covered equating approaches for alternative versions of the same test. The *Linking* section discussed other linking approaches for relating the scores and scales of different tests. In these discussions, scaling, equating, and linking activities were described and related to a wider context of testing activities, including test specifications, test form assembly, test scoring, test administration conditions (including data collection designs), and the intended purposes and uses of the test(s). When testing activities are consistent and tests are similar with respect to construction, administration, and purposes, the resulting scales are more likely to reflect their intended interpretations. When testing activities are inconsistent or are altered in tests that are scaled and linked to each other, the resulting scaling, linking, and equating procedures may not produce results that adequately support intended score interpretations.

Recent testing trends create and increase the challenges in establishing, maintaining, and relating the score scales of tests. Pursuits of alternative online administrations have increased since the COVID-19 pandemic, with mixed administration effects on reported scores (see the *Linking Tests Across Conditions of Measurement* section). There are also calls to revise testing in ways that address increasingly diverse testing populations, such as to develop and administer tests under less standardized conditions (Sireci, 2020), in ways that reflect test takers' sociocognitive backgrounds (Mislevy, 2018). In situations where testing programs compete for dwindling numbers of test takers, some programs have attempted to link their tests to other tests using substandard linking methodologies (Dorans & Moses, 2023). In other cases, unlinked scores are released with the presumption, but not necessarily the communication, that scores are linked based on nonempirical and untestable assumptions (Baldwin & Clauser, 2022; Dorans & Middleton, 2012). Assessment approaches can make different trade-offs of standardization, reliability, and comparability versus validity, group fidelity, and local uses (Brennan, 2006; Mislevy et al., 2025). Some of these trade-offs prompt recommendations to limit unwarranted score interpretations and broad comparisons of test takers that can be unfair (Dorans & Haberman, 2022; Moses, 2022, 2025).

As discussed in this chapter, large numbers of forms administered to smaller administration groups using weak data collection designs and little or no linking efforts can simultaneously inflate several types of error (random, systematic, violations of the equal construct requirement, etc.). They also raise concerns about the comparability of scores being released and used to make comparisons (Baldwin & Clauser, 2022; Moses, 2022). Experience indicates that these concerns have especially serious consequences for test score users when testing organizations are not forthcoming in communicating the procedures they use, the assumptions they make with their procedures, and the extent to which comparability in their scores is (un)supported. For these and other challenges, the recommended practices described in this chapter include checks to monitor and ensure that reported scales continue to reflect their intended interpretations. Additional recommendations are for checks on equating and linking results, especially those produced under limited and unique study conditions that may warrant interpretations that are restricted to the conditions of those linking studies rather than generalized to uses in broader testing populations (W. Lee & Brennan, 2021; Moses & Dorans, 2021; Moses, 2022).

The entity that produces reported scores has the ultimate responsibility for establishing and maintaining their reporting scales and for ensuring that the resulting scores and their interpretations are clearly communicated to test users (AERA et al., 2014). It is crucial that reported scores have clear, useful, and defensible interpretations and explanations based on sound scaling, equating, and (if necessary) linking, along with other supporting information. The ultimate goal is that reported scores, especially how they are produced and interpreted, should be integral and defensible evidence in support of validation arguments about test scores. It should be noted as well that this goal is unlikely to be attained without the active cooperation with and/or input from those who design, develop, administer, and market tests. Defensible scaling, equating, and linking is not simply the application of complex psychometrics to testing issues. All applied psychometrics involves assumptions, and the defensibility of results rests heavily on the credibility of these assumptions in the specific testing context under consideration, with appropriate attention given to quantifying and communicating likely error in reported scores:

> Whether scores are equated, linked in some weaker sense, or rescaled, however, the overarching consideration in my opinion is that users be given appropriate guidance about score interpretation and use. Part of that guidance ought to be explicit indications of the amount of error in scores and in the likely uses made of scores, as well as admonitions about likely misinterpretations of scores. (Brennan, 2007, p. 175)

ACKNOWLEDGMENTS

My chapter was greatly improved by reviews from my reviewer-collaborators, Robert Brennan, Neil Dorans, and Michael Kolen, an early review from Derek Briggs, helpful discussions with Jonathan Weeks, and suggestions from editors Linda Cook and Mary Pitoniak. I am very thankful for these reviews. I am also grateful for the collaborations with all of my coauthors cited throughout this chapter.

REFERENCES

ACT. (2001). *EXPLORE technical manual*.

Allen, N. L., Carson, J. E., & Zelenak, C. A. (1999) *The NAEP 1996 technical report*. National Center for Education Statistics.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Andrich, D. (1988). *Rasch models for measurement*. Sage Publications.

Angoff, W. H. (1962). *The equating of nonparallel tests* (ETS Research Memorandum No. RM-62-02). ETS.

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

Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement, 23*, 327–345.

Baldwin, P., & Clauser, B. (2022). Historical perspectives on score comparability issues raised by innovations in testing. *Journal of Educational Measurement, 59*, 140–160.

Becker, K. A. (2003). *History of the Stanford–Binet intelligence scales: Content and psychometrics* (Stanford–Binet intelligence scales, 5th Ed., Assessment Service Bulletin No. 1). Riverside.

Betebenner, D. W. (2009). *Growth, standards and accountability*. Colorado Department of Education.

Binet, A., & Simon, T. (1916). *The development of intelligence in children (the Binet–Simon Scale)*. Williams & Wilkins Company.

Bolt, D. (2021). Commentary on Reardon, Kalorgrides, and Ho's "Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale." *Journal of Educational and Behavioral Statistics, 46*(2), 168–172.

Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48*, 419–440.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). Academic Press.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 199–216). Springer Science + Business Media.

Brennan, R. L. (Ed.). (1989). *Methodology used in scaling the ACT assessment and P-ACT+*. CASMA.

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

Brennan, R. L. (2004). *Manual for LEGS Version 2.0*. (3). CASMA.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.), pp. 1–16). Praeger.

Brennan, R. L. (2007). Tests in transition: Discussion and synthesis. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 161–175). Springer-Verlag.

Brennan, R. L. (2010). *First-order and second-order equity in equating* (CASMA Research Report No. 30). CASMA.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes* (CASMA Monograph No. 1). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement, 50*, 204–226.

Briggs, D. C. (2022). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Routledge.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38*, 551–576.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues & Practice, 28*, 3–14.

Broussard, M. (2020, September 8). When algorithms give real students imaginary grades. *New York Times*. https://www.nytimes.com/2020/09/08/opinion/international-baccalaureate-algorithm-grades.html

Camara, W. (2020). Never let a crisis go to waste: Large-scale assessment and the response to COVID-19. *Educational Measurement: Issues and Practice, 39*, 10–18.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. Longmans, Green.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Council of Chief State School Officers.

Chen, H. H., Livingston, S. A., & Holland, P. W. (2011). Generalized equating functions for NEAT designs. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 185–200). Springer.

College Board. (n.d.). *About AP scores*. Retrieved January 1, 2025 from https://apstudents.collegeboard.org/about-ap-scores

College Board/ACT. (2018). *Guide to the 2018 ACT/SAT concordance*. https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Prentice Hall.

Cronbach, L. J. (1949). *Essentials of psychological testing*. Harper & Brothers.

CTB/McGraw–Hill. (2001). *TerraNova technical report*.

Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic b-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement, 46*, 135–158.

Davey, T., & Lee, Y.-H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised General Test* (ETS Research Report No. RR-11-26). ETS.

Davison, M. (2021). Commentary on "Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale." *Journal of Educational and Behavioral Statistics, 46*(2), 172–186.

Dorans, N. J. (1999) *Correspondences between ACT and SAT I scores* (College Board Research Report No. 99-1). College Board.

Dorans, N. J. (2002). Recentering the SAT score distributions: How and why. *Journal of Educational Measurement, 39*, 59–84.

Dorans, N. J. (2004a). Assessing the population sensitivity of equating functions. [Special issue]. *Journal of Educational Measurement, 41*(1).

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

Dorans, N. J. (2011). Holland's advice for the fourth generation of test theory: Blood tests can be contests. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 259–272). Springer.

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 20–37.

Dorans, N. J. (2013). On attempting to do what Lord said was impossible: Commentary on van der Linden's "Some conceptual issues in observed-score equating." *Journal of Educational Measurement, 50*, 304–314.

Dorans, N. J. (2018). Scores, scales, and score linking. In P. Irwing, T. Booth, & D. J. Hughs (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 573–605). Wiley Blackwell.

Dorans, N. J. (2020). Uncommon measures revisited (ETS Research Report No. RR-20-04). ETS.

Dorans, N. J. (2022, April 21–26). *Revisiting lingering linking issues: Now; Never-ending?* (Invited 2021 Robert L. Linn Distinguished Address). American Educational Research Association Annual Meeting, San Diego, CA, United States.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10, pp. 91–122). ETS.

Dorans, N. J., & Haberman, S. J. (2022). Recent challenges to maintaining score comparability: A commentary. *Journal of Educational Measurement, 59*, 251–264.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281–306.

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education, 3*(3), 245–254.

Dorans, N. J., Liang, L., & Puhan, G. (2010). *Aligning scales of certification tests* (ETS Research Report No. 10-07). ETS.

Dorans, N. J., Lin, P., Wang, W., & Yao, L. (2014). *The invariance of latent and observed linking functions in the presence of multiple latent test-taker dimensions* (ETS Research Report No. RR-14-41). ETS. http://dx.doi.org/10.1002/ets2.12041

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University, 73*, 24–34.

Dorans, N. J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement, 49*, 1–18.

Dorans, N. J., & Moses, T. (2023) Score equating: An aspirational form of score linking. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education* (vol. 14, 4th ed., pp. 236–248). Elsevier.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). Springer.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. Springer-Verlag.

Dorans, N. J., & Puhan, G. (2017). Contributions to score linking theory and practice. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment* (pp. 79–132). ETS.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 179–198). Springer-Verlag.

Dorans, N. J., & Wright, N. K. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). ETS.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and Testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Praeger.

Drasgow, F., Stark, S., Chernyshenkok, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army selection and classification decisions* (Technical Report 1311). U.S. Army Research Institute for the Behavioral and Social Sciences.

ETS. (2005). *The Comprehensive English Language Learning Assessment: Technical report*.

Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 135–159). Springer-Verlag.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61, 50–55.*

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Engelhard, G. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement, 8*, 21–38.

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments* (pp. 52–59). Routledge.

Ferguson, A., Myers, C., Bartlett, R., Banister, H., Bartlett, F. C., Brown, W., Campbell, N., Craik, K., Drever, J., Guild, J., Houstoun, R., Irwin, J., Kaye, G., Philpott, S., Richardson, L., Shaxby, J., Smith, T., Thouless, R., & Tucker, W. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science, 1*, 331–349.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). National Academies Press.

Flanagan, J. C. (1939). *The cooperative achievement tests. A bulletin reporting the basic principles and procedures used in the development of their system of scaled scores*. American Council on Education Cooperative Test Service.

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

Forsyth, R. A. (1991). Do NAEP scales yield criterion-referenced interpretations? *Educational Measurement: Issues & Practice, 10*(3), 3–9,16.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and square root. *Annals of Mathematical Statistics, 21*, 607–611.

Gao, R., He, W., & Ruan, C. (2012). *Does preequating work? An investigation into a preequated testlet-based college placement exam using postadministration data* (ETS Research Report No. RR-12-12). ETS. https://doi.org/10.1002/j.2333-8504.2012.tb02294.x

Golub-Smith, M. L., & Moses, T. P. (2014). How the scales for the GRE revised test were defined. In C. Wendler & B. Bridgeman (Eds.), *The research foundation for the GRE revised General Test: A compendium of studies* (pp. 2.2.1–2.2.9). ETS.

Gonzalez, J., & Wiberg, M. (2017). *Applying test equating methods: Using R (Methodology of Educational Measurement and Assessment)*. Springer.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Guo, H., Liu, J., Curley, E., Dorans, N., & Feigenbaum, M. (2012). *The stability of the score scale for the SAT Reasoning Test from 2005–2012* (ETS Research Report No. RR-12-15). ETS.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

Haberman, S. (2011). Using exponential families for equating. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 125–140). Springer.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40*, 254–273.

Haberman, S., & Dorans, N. J. (2011). *Sources of scale score inconsistency* (ETS Research Report No. RR-11-10). ETS.

Haberman, S. J., Sinharay, S., & Lee, Y. (2011). Statistical procedures to evaluate quality of scale anchoring (ETS Research Report No RR-11-04). ETS.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). (pp. 65–110). Praeger.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error* (Technical Report No. 15). Stanford University, Applied Mathematics and Statistics Laboratory.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). American Council on Education/Praeger.

Hanson, B. A. (1991). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design. *Journal of Educational Statistics, 16*, 93–100.

Harris, D. J. (2023). NCME presidential address: Some musings on comparable scores, *Educational Measurement: Issues and Practice, 43*, 6–15.

Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement, 10*, 35–43.

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis, 40,* 685–711.

Ho, A. D., & Haertel, E. H. (2007). *(Over)-interpreting mappings of state performance standards onto the NAEP scale* [Commissioned paper]. Council of Chief State School Officers.

Hölder, O. (1901). Die Axiome der Quantitat und die Lehre vom Mass [The axioms of quantity and the theory of measurement]. *Berichte uber die Verhandlungen der Koniglich Sachsischen Gesellschaft der Wissenshaften zu Leipzig, Mathematisch-Physische Klasse, 53,* 1–46.

Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond and Allen/Donoghue. *Proceedings of the Social Statistics Section of the American Statistical Association,* pp. 27–29.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 7–30). Springer-Verlag.

Holland, P. W. (2008, March 3–5). *The first four generations of test theory* [Paper presentation]. Association of Test Publishers Conference on Innovations in Testing, Dallas, TX, United States.

Holland, P. W. (2013). Comments on van der Linden's critique and proposal for equating. *Journal of Educational Measurement, 50,* 286–294.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Praeger.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68,* 123–149.

Holland, P. W., & Rubin, D. B. (Ed.). (1982). *Test equating*. Academic.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS Technical Report No. RR-89–84). ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–183.

Holland, P. W., & Wightman, L. E. (1982). Section pre-equating: A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. (pp. 271–297). Academic Press.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice, 3,* 8–14.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *The Iowa tests: Guide to research and development*. Riverside Publishing.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: a literature review. *PeerJ Computer Science, 5,* e208. https://doi.org/10.7717/peerj-cs.208

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion referenced interpretation. *Journal of Educational & Behavioral Statistics, 23,* 35–56.

Iowa Tests of Educational Development. (1958). *Manual for school administrators. 1958 revision.* State University of Iowa.

Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics, 10,* 143–160.

Jia, Y., & Xi, N. (2021, June 9–11). *Design considerations for linking large scale survey assessments across modes.* National Council of Measurement in Education Annual Meeting.

Jones, P., Tong, Y., Liu, J., Borglum, J., & Primoli, V. (2022). Score comparability between online proctored and in-person credentialing exams. *Journal of Educational Measurement, 59,* 180–207.

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.)*, Validation of score meaning for the next generation of assessments* (pp. 11–24). Routledge.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2,* 389–423.

Kelley, T. L. (1914). Comparable measures. *Journal of Educational Psychology, 5,* 589–595.

Kelley, T. L. (1923). *Statistical methods*. Macmillan.

Kim, H., Brennan, R. L., & Lee, W. C. (2017). Structural zeros and their implications with log-linear bivariate presmoothing under the internal anchor design. *Journal of Educational Measurement, 54*(2), 145–164.

Kim, S. Y., Kim, Y., & Moses, T. (2020). *Impact of degrees of postsmoothing on long-term equated scale score accuracy* (CASMA Research Report No. 54). CASMA. http://www.education.uiowa.edu/casma

Kim, S., & Moses, T. (2014). *An investigation of the impact of misrouting under two-stage multistage testing: A simulation study* (ETS Research Report No. RR-14-01). ETS. http://dx.doi.org/10.1002/ets2.12000.

Kim, S., & Moses, T. (2025). Item response theory (IRT) proficiency estimation methods under multistage testing. In D. Yan (Ed.), *Research for practical issues and solutions in computerized multistage testing* (Vol. 1, pp. 185–199). ETS.

Kim, S., & Walker, M. (2021). *Assessing mode effects of at home testing without a randomized trial* (ETS Research Report No. RR-21-10). ETS. https://doi.org/10.1002/ets2.12323

Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 46*(1), 36–53.

Kim, S., Walker, M. E., & McHale, F. (2010b). Investigating the effectiveness of equating designs for constructed-response tetes in large-scale assessments. *Journal of Educational Measurement, 47,* 186–201.

Kim, Y. K., Moses, T., Kolen, M. J., & Hendrickson, A. (2016). *Scaling for the SAT Suite of Assessments*. College Board.

Kim, Y. K., Moses, T., & Zhang, X. (2018a). *Student-level growth estimates for the SAT® suite of assessments*. College Board. https://collegereadiness.collegeboard.org/pdf/student-level-sat-suite-growth-estimates.pdf

Kim, Y. K., Moses, T., & Zhang, X. (2018b). *School-level growth estimates for the SAT® suite of assessments*. College Board. https://collegereadiness.collegeboard.org/pdf/sat-suite-growth-estimates.pdf

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9,* 25–44.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement, 25,* 97–110.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41,* 3–14.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Praeger.

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer Publishing.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*(4), 285–307.

Kolen, M. J., & Lee, W.-C. (2011). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1). CASMA. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.1.pdf

Kolen, M. J., & Lee, W.-C. (2012). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2). CASMA. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.2.pdf

Kolen, M. J., & Lee, W.-C. (2014). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 3). CASMA. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.3.pdf

Kolen, M. J., & Lee, W.-C. (2016). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 4). CASMA. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.4.pdf

Kolen, M. J., & Lee, W.-C. (2018). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 5). CASMA. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.5.pdf

Kolen, M. J., Tong, Y., & Brennan, R. L. (2011). Scoring and scaling educational tests. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 43–58). Springer.

Koretz, D. (2007). Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 339–353). Springer-Verlag.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement, Vol. 1: Additive and Polynomial Representations.* Academic Press.

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology, 18,* 89–109.

Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology, 64*(3), 478–497.

Lee, W., & Brennan, R. L. (2021). *An independent review of Educational Testing Service's study on predictive validity of the Graduate Record Exams for making law school admissions decisions.* https://www.americanbar.org/content/dam/aba/administrative/legal_education_and_admissions_to_the_bar/2021/20210907-casma-report-final.pdf

Lee, Y., & Haberman, S. H. (2013). Harmonic regression and cale stability. *Psychometrika, 78,* 815–829.

Lee, Y., Haberman, S., & Dorans, N. J. (2019). Use of adjustment by minimum discriminant information in linking constructed-response test scores in the absence of common items. *Journal of Educational Measurement, 56,* 452–472.

Lee, Y., & von Davier, A. A. (2011). Equating through alternative kernels. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 159–173). Springer.

Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS Research Bulletin No. RB-55-23). ETS.

Li, D. (2023). *Investigating the impact of equating on score variability using generalizability theory*. IMPS Presentation.

Lindquist, E. F. (1953). Selecting appropriate score scales for tests. Discussion. In *Proceedings of the 1952 Invitational Conference on Testing Problems* (pp. 34–40). ETS.

Lindquist, E. F. (1964, February 12–15). *Equating scores on non-parallel tests* [Paper presentation]. American Educational Research Association Annual Meeting, Chicago, IL, United States.

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

Liou, M., & Cheng, P. E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20,* 259–286.

Liou, M., Cheng, P. E., & Li, M. Y. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement, 25,* 197–207.

Liu, C., & Kolen, M. J. (2018). A comparison of strategies for smoothing parameter selection for mixed-format tests under the random groups design. *Journal of Educational Measurement, 55,* 564–581.

Liu, J., & Dorans, N. J. (2012). Assessing the practical equivalence of conversions when measurement conditions change. *Journal of Educational Measurement, 49,* 101–115.

Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice, 32*(1), 15–22.

Liu, J., Dorans, N. J., & Moses, T. (2010, May 1–3). *Evaluating the subpopulation sensitivity of the ACT–SAT concordances* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Denver, CO, United States.

Liu, J., Guo, H., & Dorans, N. J. (2014). *A comparison of raw-to-scale conversion consistency between single- and multiple-linking using a nonequivalent groups anchor test design* (Research Report No. RR-14-13). ETS.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. ETS.

Livingston, S. A. (2014). *Demographically adjusted groups for equating test scores* (ETS Research Report No. RR-14-30). ETS. https://doi.org/10.1002/ets2.12030

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3,* 73–95.

Livingston, S. A., & Kim, S. (2011). New approaches to equating with small samples. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 109–122). Springer.

Lord, F. M. (1950). *Notes on comparable scales for test scores* (ETS Research Bulletin No. RB-50-48). ETS.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics, 7,* 165–174.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48,* 233–245.

Lord, F. M. (1985). Estimating the imputed social cost of errors of measurement. *Psychometrika, 50,* 57–68.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*(2), 157–162.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 452–461.

Lorge, I. (1951). The fundamental nature of measurement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 533–559). American Council on Education.

Lottridge, S., Woolf, S., Young, M., Jafari, A., & Ormerod, C. (2023). The use of annotations to explain labels: Comparing results from a human-rater approach to a deep learning approach. *Journal of Computer Assisted Learning, 39*, 787–803.

Luce, R. D., & Tukey J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1–27.

Luecht, R., & Burke, M. (2020). Reconceptualizing items: From clones and automatic item generation to task model families. In H. Jiao & R. W. Lissitz (Eds.), *Application of artificial intelligence to assessment* (pp. 25–49). Information Age Publishing.

Marco, G. L., & Abdel-Fattah, A. A. (1991). Developing concordance tables for scores on the Enhanced ACT Assessment and the SAT. *College & University, 66*, 187–194.

McLaughlin, D. H. (2000). *Protecting state NAEP trends from changes in SD/LEP inclusion rates* (Report to the National Institute of Statistical Sciences). American Institutes for Research.

McLaughlin, D. H., & Bandeira de Mello, V. (2002, April 1–5). *Comparison of state elementary school mathematics achievement standards using NAEP 2000* [Paper presentation]. American Educational Research Association Annual Meeting, New Orleans, LA, United States.

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6*, 7–24.

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

Mislevy, R. J. (2012). Comments on Neil Dorans's NCME Career Award Address: The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice, 31*(4), 38–39.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mislevy, R. J., Olivari, M. E., Slomp, D., Wolf, A. C. E., & Elliot, N. (2025). An evidentiary-reasoning lens for socioculturally responsive assessment. In R. Bennett, L. Darling-Hammond, & A. Badrinarayan, (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. (pp. 199–241). Routledge.

Modu, C. C., & Stern, J. (1975). *The stability of the SAT score scale* (ETS Research Bulletin No. RB-75-9). ETS.

Morris, C. N. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 169–191). Academic Press.

Moses, T. (2008). Using the kernel method of test equating for estimating the standard errors of population invariance measures. *Journal of Educational and Behavioral Statistics, 33*, 137–157.

Moses, T. P. (2011). Log-linear models as smooth operators: Holland's statistical applications and their practical uses. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 185–202). Springer.

Moses, T. (2014a). Quantifying error and uncertainty reductions in scaling functions: An ITEMS module. *Educational Measurement: Issues and Practice, 33*, 29–40.

Moses, T. (2014b). Alternative smoothing and scaling strategies for weighted composite scores. *Educational and Psychological Measurement, 74*(3), 516–536.

Moses, T. (2022). Linking and comparability across conditions of measurement. Established frameworks and proposed updates. *Journal of Educational Measurement, 59*, 231–250.

Moses, T. (2025). Linking responsive assessments: Challenges and possibilities. In R. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. (pp. 180–198). Routledge.

Moses, T., Deng, W., & Zhang, Y. L. (2010). *The use of two anchors in the nonequivalent groups with anchor test (NEAT) equating* (ETS Research Report No. RR-10-23). ETS.

Moses, T., & Dorans, N. J. (2021). Aggregate-level test-scale linking: A new solution for an old problem? *Journal of Educational and Behavioral Statistics, 46*(2), 187–202.

Moses, T. P., Dorans, N. J., Miao, J., & Yon, H. (2006). *Weighting strategies for Advanced Placement exams*. ETS.

Moses, T., & Golub-Smith, M. (2011). *A scaling method that produces scale score distributions with specific skewness and kurtosis* (ETS Research Memorandum No. RM-11-04). ETS.

Moses, T. P., & Holland, P. W. (2008). *Notes on the general framework for observed score equating* (ETS Research Report No. RR-08-59). ETS.

Moses, T., & Holland, P. W. (2010). A comparison of statistical selection strategies for univariate and bivariate loglinear models. *British Journal of Mathematical and Statistical Psychology, 63*, 557–574.

Moses, T., & Kim, S. (2007). *Reliability and the nonequivalent groups with anchor test design* (ETS Research Report No. RR-07-16). ETS.

Moses, T., & Kim, S. (2015). Methods for evaluating composite reliability, classification consistency, and classification accuracy for mixed-format licensure tests. *Applied Psychological Measurement, 39*, 314–329.

Moses, T., & Kim, Y. K. (2017). Stabilizing conditional standard errors of measurement in scale score transformations. *Journal of Educational Measurement, 54*, 184–199.

Moses, T., Yang, W., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement, 44*, 157–178.

Moses, T. P., & Zhang, W. (2011). Standard errors of equating differences: Prior developments, extensions and simulations. *Journal of Educational and Behavioral Statistics, 36*, 779–803.

Moses, T., Kim, Y. K., Duffy, L., Lee, D., Patel, P., & Kulscar, B. (2021). *Linking analyses and results for digital tests in the AP 2021 administration*. [Unpublished documentation]. College Board.

Mroch, A. A., Suh, Y., Kane, M. T., & Ripkey, D. R. (2009). An evaluation of five linear equating methods for the NEAT design. *Measurement: Interdisciplinary Research and Perspectives, 7*(3–4), 174–193.

Ogasawara, H. (2001). Item response theory true score equatings and their standard errors. *Journal of Educational and Behavioral Statistics, 26*, 31–50.

Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika, 68*, 193–211.

Otis, A. S. (1922). The method for finding the correspondence between scores in two tests. *Journal of Educational Psychology, 13*, 529–545.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. National Academies Press.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*, 237–255.

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). Springer-Verlag.

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

Phillips, S. E. (2016). Legal aspects of test fairness. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement: Vol 3. NCME application of educational assessment and measurement* (pp. 239–266). Routledge.

Pommerich, M. (2007). Concordance: The good, the bad, and the ugly. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 199–216). Springer-Verlag.

Pommerich, M. (2016). The fairness of comparing test scores across different tests or modes of administration. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement: Vol 3. NCME applications of educational assessment and measurement* (pp. 111–134). Routledge.

Puhan, G., & Kim, S. (2022). Score comparability issues with at-home testing and how to address them. *Journal of Educational Measurement, 59*, 161–179.

Puhan, G., Moses, T., Grant, M., & McHale, F. (2009). Small-sample equating using a single-group nearly equivalent test (SiGNET) design. *Journal of Educational Measurement, 46*, 344–362.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute Educational Research.

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district distributions to a common scale. *Journal of Educational and Behavioral Statistics, 46*, 138–167.

Rijmen, F., Manalo, J., & von Davier, A. A. (2009). Asymptotic and sampling-based standard errors for two population invariance measures in the linear equating case. *Applied Psychological Measurement, 33*, 222–237.

Schalet, B. D., Lim, S., Cella, D., & Choi, S. W. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika, 86*(3), 717–746.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 181–198). American Psychological Association.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105.

Sinharay, S., & Holland, P. W. (2006a). *Choice of anchor test in equating* (ETS Research Report No. RR-06-35). ETS.

Sinharay, S., & Holland, P. W. (2006b). *The correlation between the scores of a test and an anchor test* (ETS Research Report No. RR-06-04). ETS.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184–203.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 1–76). John Wiley.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36*(4), 336–346.

Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329–346.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287–310). Springer-Verlag.

Thissen, D. (2016). Chapter 11: Commentary on the assessment of the fairness of comparisons under divergent measurement conditions. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 203–214). Routledge.

Thissen, D., Liu, Y., Magnus, B., & Quinn, H. (2015). Extending the use of multidimensional IRT calibration as projection: Many-to-one linking and linear computation of projected scores. In L. van der Ark, D. Bolt, W. C. Wang, J. Douglas, & S. M. Chow (Eds.), *Quantitative psychology research: Springer Proceedings in Mathematics & Statistics* (Vol. 140, pp. 1–16). Springer.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Lawrence Erlbaum Associates.

Thorndike, E. L. (1910). The measurement of the quality of handwriting. *Teachers College Record, 11*, 86–151.

Thorndike, E. L. (1919). *An introduction to the theory of mental and social measurements.* Columbia University, Teachers College.

Thorndike, E. L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology, 6*, 29–33.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*(7), 433–451.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology, 17*, 446–457.

Thurstone, L. L. (1927). The law of comparative judgement. *Psychological Review, 34*, 278–286.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs. No. 1*. University of Chicago Press.

van der Linden, W. J. (2011). Local observed-score equating. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 201–223). Springer.

van der Linden, W. J. (2013). Some conceptual issues in observed-score equating. *Journal of Educational Measurement, 50*, 249–285.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer.

von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking*. Springer.

von Davier, A. A. (2021). Commentary on "Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale." *Journal of Educational and Behavioral Statistics, 46*(2), 203–208.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). *The kernel method of test equating*. Springer Publishing.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). The chain and post-stratification methods of observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement, 41*, 15–32.

von Davier, A. A., & Liu, M. (2007). Population invariance. [Special issue]. *Applied Psychological Measurement, 32*(1).

Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics, 44*, 390–414.

Wang, J., & Smith, R. (2003). *NYSESLAT calibration and linking summary*. ETS.

Wang, T. (2011). An alternative continuization method: The continuized log-linear method. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 141–157). Springer.

Wang, T., & Brennan, R. L. (2008). A modified frequency estimation equating method for the common-item nonequivalent groups design. *Applied Psychological Measurement, 33*, 118–132.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *Journal of Educational Measurement, 38*, 19–49.

Weeks, J. (2018). An application of multidimensional vertical scaling. *Measurement: Interdisciplinary Research and Perspectives, 16*, 139–154.

Wendler, C., & Bridgeman, B. (2014). *The Research Foundation for the GRE (R) revised General Test: A compendium of studies*. ETS.

Williams, V. S. L., Rosa, K. R., McLeod, L. D., Thissen, D., & Sanford, E. E. (1998). Projecting to the NAEP scale: Results from the North Carolina end-of-grade testing program. *Journal of Educational Measurement, 35*, 277–296.

Wise, L. L. (2017). Commentary I: Validation of score meaning in the next generation of assessments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 52–59). Routledge.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97–116.

Wright, B. D. (1994). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions, 6*, 196–200.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–283). Springer-Verlag.

Yen, W. M., & Fitzpatrick, A. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Praeger.

Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement, 28*, 274–289.

Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests*. Holt.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Springer.

Zesch, T., Horbach, A., & Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement: Issues and Practice, 42*, 44–58.

NOTE

1. The variance of mean-score differences plus the relative error variance in classical test theory is the absolute error variance in generalizability theory (Brennan, 2001).