CCSSO
Council of Chief State School Officers

REVISED FIRST EDITION

A Practitioner's Introduction to Equating

with Primers on Classical Test Theory and Item Response Theory

Prepared in collaboration and with the support of
Technical Issues in Large Scale Assessment (TILSA)
State Collaborative on Assessment and Student Standards (SCASS)
of the Council of Chief State School Officers (CCSSO)

by

Joseph Ryan, Arizona State University
Frank Brockmann, Center Point Assessment Solutions

The Council of Chief State School Officers

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

State Collaborative on Assessment and Student Standards

The State Collaborative on Assessment and Student Standards (SCASS) projects were created in 1991 by the Council of Chief State School Officers to encourage and assist states in working collaboratively on assessment design and development for a variety of topics and subject areas. These projects are organized and facilitated within CCSSO by the Division of State Services and Technical Assistance.

Technical Issues in Large-Scale Assessment (TILSA)

TILSA is part of the State Collaborative on Assessment and Student Standards (SCASS) project whose mission is to provide leadership, advocacy, and service by focusing on critical research in the design, development, and implementation of standards-based assessment systems that measure the achievement of all students. TILSA addresses state needs for technical information about large-scale assessment by providing structured opportunities for members to share expertise, learn from each other, and network on technical issues related to valid, reliable, and fair assessment; designing and carrying out research that reflects common needs across states; arranging presentations by experts in the field of educational measurement on current issues affecting the implementation and development of state programs; and developing handbooks and guidelines for implementing various aspects of standards-based assessment systems.

This long-standing partnership has conducted a wide variety of research over the years into critical issues affecting the technically-sound administration of K–12 assessments, including research on equating, setting cut-scores, consequential validity, generalizability theory, use of multiple measures, alignment of assessments to standards, accommodations, assessing English language learners, and the reliability of aggregate data. In addition, TILSA has provided professional development in critical topics in measurement for its members. The partnership has developed technical reports on each topic it researched. In addition, TILSA has produced guidelines for developing state assessment programs.

The Council of Chief State School Officers
Christopher Koch, President
Gene Wilhoit, Executive Director

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000
Fax (202) 408-8072
www.ccsso.org

Table of Contents

Acknowledgements

This publication was sponsored by the Council of Chief State School Officers (CCSSO) and developed in cooperation with the Technical Issues in Large-Scale Assessment (TILSA) collaborative under the leadership of Doug Rindone, who played a critical role in advancing this publication and supporting its goals. TILSA's Subcommittee on Equating initiated the project, led by Subcommittee Chair Michael Muenks (Missouri Dept. of Elementary and Secondary Education). Duncan MacQuarrie (CCSSO) and Phoebe Winter (consultant) provided leadership and guidance throughout the development of this document. Joseph Ryan (professor emeritus, Arizona State University) served as chief editor and provided research, writing, editing and psychometric/equating expertise. Frank Brockmann (Center Point Assessment Solutions) drafted much of the original text, composed diagrams, and provided ongoing logistical and project management support.

The TILSA collaborative provided direct input to help shape the development and revision of this document. The TILSA Subcommittee on Equating reviewed draft material in June 2007 (Nashville, TN), October 2007 (Salt Lake City, UT), February 2008 (Atlanta, GA), and June 2008 (Orlando, FL), and the TILSA group as a whole also provided valuable feedback during these meetings that helped guide the project as it moved forward.

During its later stages of development, this handbook was reviewed externally by the Technical Special Interest Group of National Assessment of Educational Progress (NAEP) coordinators. Comments from this external review helped provide additional editorial refinement and focus for this document. Contributing members include:

NAEP State Coordinators
Vickie Baker, West Virginia
Kate Beattie, Minnesota
Barbara Bianchi, New Mexico
Pauline Bjornson, North Dakota
Challis Breithaupt, Maryland
Dianne Chadwick, Iowa
Mike Chapman, Montana
Mark Decandia, Kentucky
William Donkersgood, Wyoming
Jeanne Foy, Alaska
David Gebhardt, New Hampshire
Wendy Geiger, Virginia
Carrie Giovanonne, Arizona
Cynthia Hollis, Missouri
Patsy Kersteter, Delaware
Tor Loring-Meier, Nevada
Jo Ann Malone, Mississippi
Angie Mangiantini, Washington
Jan Martin, South Dakota
Andy Metcalf, Illinois
John Moon, Nebraska
Pam Sandoval, Colorado
Renee Savoie, Connecticut
Barbara Smey-Richman, New Jersey
Bert Stoneberg, Idaho
Jessica Valdez, California

NAEP State Coordinators (Former or Interim)
Therese Carr, South Carolina
Elaine Hultengren, Oregon
Chris Webster, South Carolina
Carolyn Trombe, New York

Tribal Urban District Assessment Coordinators
Margaret Bartz, Chicago Public Schools
Maria Lourdes de Hoyos, Austin Independent School District

State Department of Education, Bureau of Assessment
Gail Pagano, Connecticut

NAEP Coaches
Dale Carlson
Gordon Ensign
James Friedebach
Carole White

NAEP State Service Center
Jason Nicholas

Grateful thanks also to Michael Kolen, University of Iowa, for his critical review and very thoughtful suggestions for this handbook and to Hariharan Swaminathan, University of Connecticut, who offered advice on an earlier version of the text.

Introduction

1. Audience

This handbook focuses primarily on equating test forms.

Equating is a technical procedure or process conducted to establish comparable scores on different versions of a test, allowing them to be used interchangeably. It is an important aspect of establishing and maintaining the technical quality of a testing program by directly impacting the *validity* of assessments—the degree to which evidence and theory support the interpretations of test scores. When two test forms have been successfully equated, educators can validly interpret performance on one test form as having the same substantive meaning compared to the equated score of the other test form.

There are a number of substantive and technical issues involved in equating and many potential pitfalls in its use. This handbook was written for decision makers to guide them in addressing these issues and to help them avoid potential problems. Intended as both a guide and teaching tool, it aims to provide readers with the practical knowledge needed to make appropriate decisions, especially readers who may have arrived at their current position from a non-technical background.

Therefore, this publication is for

- newly appointed assessment personnel coming from non-technical disciplines who need practical guidance with regard to equating decisions and their potential impacts
- experienced psychometric experts who may benefit by offering an equating primer as a resource to non-psychometrician colleagues to encourage a better understanding of the issues
- policy personnel in the position of explaining the reasoning behind prior equating decisions or advocating the future direction of an assessment program
- psychometricians who would benefit from basic models that illustrate past decisions and how they related to policy
- anyone who might benefit from a better understanding of what equating is, why it is done, and how common problems might be avoided

2. Purpose

Equating is an essential tool in educational assessment due the critical role it plays in several key areas: establishing validity across forms and years; fairness; test security; and, increasingly, continuity in programs that release items or require ongoing development.

Although the practice of equating is rooted in long standing practices that go back many decades, one of the driving forces behind this handbook has been the notion that a great deal of information about equating is unfamiliar to or not easily accessed by practitioners. This information appears in scholarly texts, professional journals, and other publications. It derives from the accumulated experiences of people who are trained and experienced in measurement and have highly technical understandings of the issues; however, even when expert consultants advise policymakers on equating-related matters, communicating the substantive significance of the technical concepts in a user-friendly manner is a challenge.

Thus the primary aim of this handbook is to provide an abbreviated conceptual background of key concepts and describe some common equating issues. It also attempts to guide readers to taking the first steps toward viable solutions. This handbook is **not** an attempt to gather or present a collection of mathematical equations, charts, graphs, or statistical formulae as part of any technical analysis. In fact, the use of equations to explain concepts has been deliberately minimized and avoided whenever possible.

The handbook's secondary aim is to provide state-level measurement professionals with a useful resource for communicating the rationale for equating decisions to policymakers and other stakeholders. This handbook can begin to bridge the gaps between various perspectives as they relate to equating by describing situations where psychometric and policy matters might interact. As such, it may be a starting point or guide for formulating solutions that are practical, feasible, and technically sound.

3. Industry Standards and Other References

Readers of this handbook are strongly encouraged to refer to *Standards for Educational and Psychological Testing* (1999), a publication of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The standards were developed to "represent the current consensus among recognized professionals regarding expected measurement practice" (p. viii). They are intended "to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices" (p. 1). The standards are the most widely recognized collection of documented principles and practices of the measurement community.

This handbook draws upon information from a variety of resources about scaling, linking, equating, and other core concepts of classical and modern test theory. These resources are listed at the end of the handbook. The handbook also draws upon the experience of measurement professionals who have encountered situations in which a deep understanding of both the technical and conceptual aspects of equating was required to solve problems. Their experiences in the field provide the basis for the "guiding" aspect of the handbook (see Acknowledgements).

4. How This Handbook Is Organized

This handbook contains five chapters. Readers do not need to progress through each chapter sequentially. However, because many equating concepts are interrelated, reviewing chapters 1, 2, and 3 first is helpful. This is especially true for readers who have modest prior training in the technical aspects of educational measurement.

Each chapter of the handbook has a combination of teaching and guiding objectives. The background section (chapters 1 and 2) provides a foundation for understanding basic measurement concepts for people new to assessment. The equating designs and procedures section (chapters 3 and 4) helps readers see how the background chapters apply to the equating designs and procedures most commonly seen in practice. The last chapter aims to guide readers toward better-informed decisions by describing common equating-related issues and provides a set of questions to ask test contractors or research psychometricians. This chapter reflects the practical experience of measurement professionals in the field and will help both new and experienced readers.

Each chapter is briefly described here:

Background

Chapter 1: Overview
Introduces equating and explains why it matters to large-scale testing programs—statewide programs in particular. It describes the relationship between linking and equating and how they differ. This part also covers the most common misconceptions about equating.

Chapter 2: A Classical and Modern Test Theory Primer
Examines these two basic approaches to measurement and includes a very basic primer. Terminology is explained (IRT, item parameters, et cetera), and its proper use is explicated.

Equating Basics

Chapter 3: Equating Designs
Illustrates the fundamental features of the most commonly used equating designs.

Chapter 4: Equating in Practice
Covers the "how" of the equating process and its basic procedures. This chapter explains the basic concepts people need to understand to make appropriate decisions; it is intended to help both new and experienced practitioners discuss technical matters with testing contractors and other measurement professionals.

Guiding

Chapter 5: Common Equating Issues
This part covers current topics related equating and includes a set of guiding questions for each topic.

5. Conventions Used in This Handbook

The handbook is organized to help readers acquire a basic understanding of equating in a relatively short amount of time. It uses several conventions to illustrate key ideas and help readers navigate the material.

Conceptual Overviews

Key sections of the handbook provide conceptual overviews. These overviews provide the essential concepts needed to help orient the reader before following up with detailed technical information.

Conceptual Diagrams

Simple diagrams are used throughout the handbook to visually represent key ideas or concepts. These figures are not intended to be detailed technical schematics.

Minimal Equations

The handbook makes minimal use of mathematical equations and highly technical descriptions. Instead, visual representations or verbal examples are provided whenever possible.

Chapter Glossaries

Words shown in bold text when first used are defined in the glossary at the end of the chapter in which they appear. Italics are used for emphasis.

References

References are provided at the end of the handbook for readers wanting additional resources and more technical information.

Symbols

All sections of the handbook use icon symbols to call attention to key concepts or questions.

Symbols/Icons Used

This symbol highlights areas where background information or key ideas are provided to help the reader along.

This text symbol calls attention to or emphasizes key ideas.

This text symbol calls attention to common questions.

This symbol acts as a "caution" or "warning" sign to highlight practices that may lead to trouble down the road.

This symbol calls attention to cases for which a more detailed idea or description is offered.

This symbol calls attention to "rules of thumb."

This symbol calls attention to descriptions of equating terms that are useful or necessary for a better understanding of equating concepts.

Chapter 1

An Overview of Assessment, Linking, and Equating Concepts

This overview introduces the concept of equating and explains why equating is important to most large-scale testing programs (and statewide programs in particular). It also describes the relationship between linking and equating and covers the most common misconceptions about equating.

1-A. Valid Inferences about Students: The Purpose of Assessment

Modern educational assessment programs are used for a variety of purposes: to improve student learning of content standards through improved instruction based on the assessment results; to complement curriculum or teaching methods; to inform teachers/students of their progress; to inform the public about school performance; to be used as a guide in decision making about students, teachers, or schools; and to provide various data comparisons (Redfield, 2001, p. 8).

The information that serves these purposes is derived from *individual tests* that make up an assessment *program*. The purpose of any particular assessment, however, is more specific—that is, its purpose is to give users an accurate description of what students know and are able to do.

On most large-scale tests, students respond to a set of test items intended to represent the entire **domain** of all possible items or tasks. For example, in early grades we might want to know if students can do one- and two-digit addition. We then give students a test form with a sample (perhaps 20 items) of all such questions. Depending on how they perform, we can make valid inferences about the degree of their knowledge about the entire learning domain of one- and two-digit addition (a domain with some 10,000 tasks).

The most important characteristic of any assessment procedure is its impact on **validity**. *Standards for Educational and Psychological Measurement* defines validity as "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests" (AERA, APA, and NCME, 1999, p. 9).

1-B. Validity and the Industry Standards

The standards also state that "[v]alidation logically begins with an explicit statement of the proposed interpretation of test scores" and that such an interpretation "refers to the construct or concepts the test is intended to measure" (AERA, et al., p. 9).

In other words, we need a clear explanation of what students are intended to know and be able to do in order to determine the soundness of any educational assessment. If we do not know explicitly what students are intended to learn and be able to do, we cannot evaluate how well (validly) an assessment procedure works in measuring student learning.

Linn (2008) further clarifies by stating that although casual discussions often refer to the validity of an assessment, it is the uses, interpretations, and claims about assessment results that are validated, not test results themselves:

> Evidence may support the conclusion that a particular use of assessment results has good validity. That same assessment may produce results, however, that have little or no validity when interpreted or used in a different way. For example, an assessment may provide a good indication of what students know and can do in a specified content area and provide information that is useful in instructional planning, but have inadequate validity for making high-stakes decisions about individual students such as the award of a high school diploma. (p. 1)

1-C. Learning Targets

The practice of describing learning targets is a discipline unto itself. Consider the following descriptions:

1. The student will learn about Sir Isaac Newton.
2. The student will describe the three major tenets of Newton's laws of motion.

In this highly simplified example, the second description clearly provides a better basis for assessment than the first. The second description specifies what is expected of students and therefore provides a framework for evaluating how well the assessment supports inferences about intended student learning.

Many frameworks exist that can help to systematically describe what we might expect students to know and do. Benjamin Bloom's *Taxonomy of Educational Objectives* (1956) is a traditional resource. In this framework, learning is defined in terms of the cognitive processes of comprehension, application, analysis, synthesis, and evaluation. Bloom's work is also part of a handbook for practitioners of classroom assessments (Bloom, Hastings, & Madaus, 1971), and *Taxonomy* was later revised and expanded upon (Anderson & Krathwohl, 2001). Other influential sources have provided their own frameworks for classifying learning, including the National Assessment of Educational Progress (NAEP), Robert Marzano and John Kendall (2007), and Norman Webb (1997); each framework focuses on a particular set of learning dimensions pertaining to cognitive processes or content knowledge.

The most critical consideration is not *which* framework is employed, but that some clear, deliberate, and well-understood method for describing learning goals and targets is in place. Whatever framework is used should be easily recognized and familiar to all participants within the educational program. The validity of an assessment program cannot be evaluated without this kind of clear focus or target.

The discussion of learning targets is not always included when considering linking and equating. However, the adequacy of equating should be judged in terms of the validity of inferences that scores on equated test forms have the same substantive meaning with respect to what students are expected to know and be able to do. Learning targets define and explain the expectations and are therefore central to the interpretation of test results.

The mode or format of the assessment is an additional dimension/description beyond the content and cognitive processes used in **test specifications** and other documents. Mode or format information is added so that the descriptions pertain to content, cognitive process, and assessment format. Formats can include **selected response** items (e.g., multiple-choice, true-false, matching) and **constructed response** items (e.g., short answer, extended responses, work samples).

1-D. Linking and Equating

Most large-scale assessment programs utilize more than one test form. Thus, successful **equating** is an important factor in evaluating assessment validity. Equating is a technical procedure or process conducted to establish comparable scores, with equivalent meaning, on different versions of test forms of the same test; it allows them to be used interchangeably. As such, it often becomes an important topic of discussion within testing programs. However, the term *equating* is often misunderstood or used inappropriately. Sometimes, the terms equating and linking are used synonymously as general terms, but this can be misleading.

Linking is the practice of pairing or matching scores on two test forms with no strong claim that the paired scores have the same substantive meaning. Linking is a concept different from equating and does not support the same interpretations supported by equating. Some of the confusion in the use of these terms is likely based on the fact that the same *procedures* are used in both linking and equating.

In both linking and equating, the scores on one test form are matched to or paired with scores on another test form. For example, students' scores on a statewide **standards-based assessment** (SBA) can be paired or linked to scores on a standardized **norm-referenced test** (NRT). Such a linking would result in a table with two columns; each row would link a score on a state test form to a particular score on a nationally norm-referenced test form (and vice-versa), thereby linking the two as shown in Figure 1.0:

Standards-Based Assessment (SBA) Score	Norm-Referenced Test (NRT) Score
325	422
333	429
341	437

Figure 1.0

The proper interpretation of this linking is described with phrases like

- "Students who make a 325 on the SBA will most likely earn a score of 422 on the NRT."
- "Kids who get a 437 on the NRT would be expected to score a 341 on the SBA."

Equating two test forms supports a much stronger claim. If the SBA and NRT were successfully equated, a valid interpretation would be

- "Students with a score of 325 and students with an NRT score of 422 have a very similar level of knowledge and skill with respect to what is being measured."

To say two forms are *equated* is to say that they measure the same content and cognitive processes and support the same inferences about what students know and can do. This is a *very* strong claim.

Linking is a much weaker claim that merely asserts an association between scores on different assessments. Equating also asserts an association between scores, but equating has the additional connotation that these paired scores have the same substantive meaning.

IMPORTANT

In the example shown in Figure 1.0, the score on the SBA cannot *automatically* be substituted for the NRT score *as though it were* the NRT score. Linking is simply a process for empirically pairing two scores with no claim that the paired scores have the same substantive or technical meaning.

The most common application of formal equating in statewide assessment programs occurs where test forms are equated within a given grade level from one year to the next. In these situations, equated scores from a grade 5 reading assessment, for example, have the same meaning in 2006 as in 2007.

Given these distinctions, linking and equating can be better understood by thinking about equating as one end of a linking continuum, as shown in Figure 1.1:

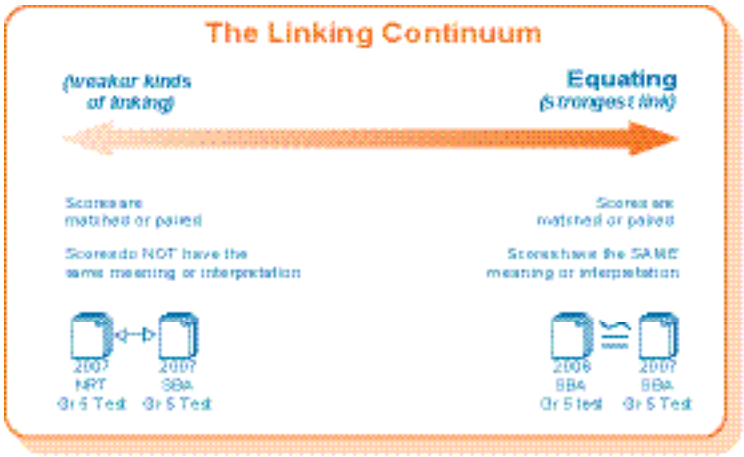


Figure 1.1: The Linking Continuum

There are cases in which carefully planned efforts are made to construct and then equate two different versions or forms of a test. In other cases it is clear that the scores on two different test forms are being linked or paired, but no strong claims are made that scores on one test have the same meaning as scores on the other.

Between the ends of the linking continuum, there are assessment situations that approximate equating in the strict sense but do not meet all the requirements for equating. These situations fall somewhere along the range of the continuum. By thinking about equating this way—as points along the range of a linking continuum—we can begin to clarify the language we use to describe a particular situation.

In many cases, determining the equivalence of two test forms requires evaluating a collection of substantive and technical information and making a professional judgment. For example, over time it is common for states to modify their content standards and modify the design of the state assessment to reflect the revised curriculum expectations (which supports the validity of the test). Such changes might involve adding or removing certain learning targets and the items that measure them, or changing the number of items or points used to assess certain subskills. In other cases, states sometimes change the assessment format and use a different balance of multiple choice and **constructed response** items. Just how much change can be made without violating "equivalent meaning" across forms is not clear. Various technical procedures can provide information that is useful in deciding whether two forms are equivalent, but professional judgment and curricular and technical expertise is required to make the final call.

In contrast to *linking*, equating supports the claim that a student who earns a given score in 2006 knows and performs similarly to a student who earns the equivalent score in 2007. This is especially important in *maintaining equivalent meaning of cut scores and performance levels* from one year to the next.

Consider the following questions:

Question 1: "What educational activity is being considered that would require (or be supported by) linking or equating test forms?"

Now consider the following measurement activities in response to the question above:

Activity 1: "We are creating new test forms and we need to equate the old and new forms within each grade level tested."

Activity 2: "We're going to build a vertical scale to equate each grade's test forms to all other test forms for grades 3 through 12." (See glossary for **vertical** scale.)

Activity 3: "We want to find out how scores on an external assessment (a commercially available norm-referenced test) could be equated to a score on our state test. That would tell us which score on the [nationally normed test] is most likely or most often earned by students who are at or above the 'proficient' cut score on our state test."

Even though each answer refers to "equating" in a generic sense, the last two responses cannot support the claim that the test forms to be equated necessarily measure the same construct and/or should have identical meaning. If we think about these scenarios along the range of linking continuum, the second and third scenarios seem better described as situations that are toward the **"weaker" linking** end. (We cannot be certain about the first answer without some additional information). Thus, another question that might help to place these kinds of scenarios on the linking/equating continuum is:

Question 2: "Is it reasonable to imagine the tests involved actually measure similar content in a way that supports equated scores—that is, scores that *mean the same thing* and can be used *interchangeably*? If not, why not?"

To say that two test forms are truly equated is a strong assertion. In many cases, people may *intend* to equate two test forms only to find out after the fact that conditions did not exist to support inferences that can be made from equated scores. In such cases, we might assert that the scores have been linked, but the claim that the scores are equated would be rejected.

1-E. Common Misconceptions about Equating

It is critical to note that the process of equating test forms begins with the very design and construction of those test forms. Two test forms that are to be equated should be constructed to measure the same content and cognitive processes, and they should use the same test question formats as well. It is still possible in some situations to equate forms that are not constructed to be completely parallel in content and format, but similarly constructed test forms are more likely to yield useful results.

The term *equating* is often misunderstood or used inappropriately. The misconceptions and mistaken beliefs about what equating is—and is not—are part of the motivation behind this handbook. Such misconceptions usually break down into several types: equating as a threat, equating as a shortcut, equating as repair shop, semantic misappropriations, and lastly, no concept of equating at all—seeing equating as nothing more than a mystery.

Equating as a Threat to Measuring Gains
Not uncommon is the misconception that if a new test form is equated to a prior year's test form, the equating process will wipe out hard won improvements in test scores. The thinking goes something like this: State Test Form A is created, then administered in 2007. After schools work hard to improve results, scores on State Test Form B improve in 2008—and the gains are recognized, admired, or perhaps even celebrated. If a new test selected for 2009 (Form C) is equated to the original 2007 exam, it will have to be made "harder" in order to "equal" the schools' prior 2007 performance, thereby wiping out the gains seen in 2008.

The basis for this misconception lies in the notion that equated test forms are test forms of equal difficulty. This is not the case—rather, equating takes into account relative *differences* in the difficulty of the equated test forms in estimating the achievement of students. Thus, in the example illustrated in Figure 1.1, equating the 2009 test form with the 2007 test form would actually *preserve* any real 2008 gains by holding the measurement scale constant. Students' gains would be evident precisely because the measurement scale was not allowed to change, even if equated test forms did not fall in exactly the same position on the scale.

Equating as a Tool for Universal Applications
Sometimes the allure of test linking and equating can give rise to the notion that almost any test form can be equated to some other assessment. For example, if the same examinees are taking both a reading test and a writing test, couldn't their two test forms be linked in such a way that future examinees would only need to take one test or the other for us to know how they would score on *both* tests? Although the motivation for using these kinds of links can be compelling (e.g., reduction in testing time, cost reductions, less administrative overhead), measurement professionals will often point to the reliability and validity issues involved in such linkages and the claims they imply. The potential for less-than-valid inferences depend on how the results will be used; for example, interpreting a score on a *reading* test as though it has the same meaning as a score on the *writing* test is not generally considered a valid interpretation.

Equating as a Repair Shop
This misconception refers to the belief that by equating test forms, problems rooted in test development can be corrected. In this erroneous view, items used operationally that are later found to be problematic, based on substantive technical review, can be "equated away."

People new to assessment sometimes see equating as a sort of mathematical equalizer tool capable of absorbing a multitude of variations between two test forms: significant changes in item positioning, changes to the content standards that the items are

intended to measure, and changes to the items themselves. In fact, changes such as these are not factored into the equating but instead pose real challenges—and sometimes outright threats—to validity.

Semantic Misappropriation
Unlike the nuanced differences in the types of linking/equating along the linking continuum, semantic misappropriation refers to any use of the term equating that sounds plausible but may be incorrect. For example, using the term *equating* in place of *alignment* is a semantic misappropriation; to the layperson, the verbs *equating* and *aligning* may seem similar—especially in the context of testing and test items. But in the measurement community, *alignment* refers to the degree to which a test and its test items are in concert with stated learning goals or specifications and is an activity quite distinct from equating.

Equating as a Mystery
This misconception is actually the absence of any conception of equating whatsoever. Some people may be entirely unfamiliar with the term or its principles and practices. Those coming to the measurement community from unrelated fields may have only a vague notion about equating—such as the fact that test items are "statistically analyzed" in order to "make sure tests are equally easy or difficult."

Concluding Thoughts about the Fundamental Concepts of Equating

There is no simple "cookie cutter" outline for making decisions about equating test forms. Certain guiding questions are very useful in framing the issues that should be considered, and answers to these questions often lead the way to an appropriate equating design. Among the key questions are

What is the purpose of this assessment?

- What kinds of inferences need to be made from the resulting information?
- What future decisions will need to be made as a result?
- What resources can be made available to support the process?
- Which students, schools or school districts will participate?
- Will the form used to collect equating data look like the operational test form?
- How closely can the test sessions used to collect equating data resemble the actual test administration?
- Will students, teachers and test administrators know that they are involved in a testing equating data collection that does not have the same stakes or consequences as the real test?

Each year, many large-scale assessments are administered throughout the United States in such a way that test takers view testing day as "taking *the* test" and not "taking *a particular form of* the test." When users of the test consider results of Form A to be the same as those from Form B, this is a highly desired result.

According to psychometric experts, "equating is successful to the extent that the form taken is a matter of indifference to each examinee" (Kolen & Brennan, 2004, p. 430). In other words "Statewide Test X" is successfully equated to the degree that examinees' performance does not depend in any way on whether they are handed Form A or Form B on test day. Why? Because equated test forms are considered *interchangeable*, and examinees can be expected to get equivalent scores regardless of which form each individual takes.

Chapter Glossary:

Construct
The underlying theoretical concept or characteristic a test is designed to measure.

Constructed-response
Items that require students to create their own responses or products rather than selected-response where students choose a response from enumerated sets.

Domain
The set or collection of all knowledge elements and skills in a subject matter area deemed important for teachers to teach and students to learn.

Equating
The practice of placing two or more tests on the same scale and satisfying other requirements in order to use test scores *interchangeably*, as having the same meaning.

Linking
The practice of matching or pairing scores on one test to scores on another test. Linking refers to the process of connecting scores from different tests for the purpose of predicting, comparing, or (if certain conditions are met) using results interchangeably (e.g., they are equated). There is no claim that the linked scores necessarily reflect the same learning, knowledge, or skills.

Norm-Referenced Test (NRT)
A test whose interpretations and scores are based on a comparison of a test taker's performance to the performance of other people in a specified reference population.

Psychometrics
Psychometrics is defined as the branch of psychology that deals with the design, administration, and interpretation of quantitative tests for the measurement of psychological variables. Psychometricians are the practitioners of this field.

Raw Score
Sometimes defined as the number of items answered correctly or points earned. In multiple-choice test results, the raw score is typically expressed as the number of points earned, without any regard for position in relation to other students. For constructed-response questions or other question types that require some form of judgment to score, the raw score is typically expressed as the total number of score points earned overall.

Selected-response
A test item that requires students to select an answer from a list of given options. Common selected-response formats include multiple-choice, true-false, and matching.

Scaling
The process of associating numbers (or other ordered indicators) with the performance of individual test takers. **Raw scores** are transformed to percentages by dividing by the number of points attainable; raw scores are converted to *scale scores* using statistical methods. Typically, *scales* are constructed in ways that will help test users interpret the scores.

Standards-Based Assessment (SBA)
An assessment system that assesses student performance at different grade levels based on publicly adopted standards of what is to be taught is a Standards-Based Assessment. A standards-based assessment system is designed to hold schools publicly accountable for each student's meeting those high standards. Often, standards-based assessment systems have different levels of achievement that define performance categories.

Test Form
A collection of test questions or tasks assembled, published, and administered to examinees. Each form is typically labeled under a versioning scheme (e.g., 2008 Form C or Form G4-B) to identify it as one of several versions of a test that are considered interchangeable: they measure the same *constructs*, are intended for the same purposes, and are administered using the same directions.

Test Specifications
The frameworks that specify the proportion of items that assess each content and process/skill area as well as the format of items, responses, and scoring protocols and procedures. These frameworks also specify the desired psychometric properties of the test and test items, such as the distribution of item difficulty.

Transformation
The process of converting raw scores to scale scores. Raw scores are *transformed* to scale scores using various statistical calculations and methodologies. Reporting a percentage correct is a transformation of a raw score.

Validity
An overall evaluation of the degree to which accumulated evidence and theory support specific interpretations of test scores. The appropriateness of the inferences that can be made on the basis of test results.

Vertical Scale
A common scale that includes test scores from multiple grades or difficulty levels within a subject area, typically constructed to track student progress over time (see Chapter 5).

Chapter 2

A Primer of Classical and IRT Measurement Theories

This section of the handbook provides a primer of two key theories/approaches to measurement that guide nearly all large-scale educational assessments: Classical Test Theory and Item Response Theory.

The material presented in Chapter 2 is intended to communicate the core concepts of these theories in a highly conceptual way. This chapter provides a brief discussion of these theories and explains why they are used. Readers with training in educational measurement, especially those who have received technical training, will be familiar with these concepts and may wish to scan or skip this material.

Measurement professionals generally agree that all members of the assessment community—especially those who are not trained in measurement science—stand to benefit from a better understanding of the basic principles of the two primary approaches to measurement: **Classical Test Theory (CTT)** and **Item Response Theory (IRT).**

CTT and IRT literacy—even in terms of the basic conceptual frameworks—has not kept pace with usage. As a result, policymakers and other professionals within the assessment community are often placed in positions of influence or authority without a solid conceptual understanding of the core concepts. This handbook was written in direct response to these situations, and this section of the handbook provides a highly conceptual primer for these theories.

For many reasons, the use of IRT is now widespread and almost all statewide programs employ it along with CTT. IRT has numerous implications for equating and test construction. IRT's popularity among statewide assessment and accountability programs may be the result of its ability to "bypass" some inherent limitations of CTT in certain situations. This section begins with CTT as a starting point.

2-A. Fundamental Concepts of Classical Test Theory

CTT refers to a body of knowledge that emerged from statistical measurement approaches used since the early 1900s. In sharp contrast to IRT, the earlier CTT approaches focused on observed raw scores.

The fundamental model of CTT is that observed raw scores are composed of two components: the "true" score the person would make if measurement were perfect; and the "error" that might reflect shortcomings of the items or test, idiosyncrasies of the particular testing setting, or variation in the students' ability to perform.

Under CTT, the core assumption underlying true scores is that any score an examinee receives on a test (the observed score) comprises two hypothetical components: an examinee's **true score**, and some amount of random **error** (Crocker & Algina, p. 107). This fundamental model is often expressed symbolically as

$$O = T + E$$
(Observed score = True score + Error)

In this equation, O represents the examinee's observed score on a test, T represents the examinee's true ability ("true score"), and E represents random error. In practice, the "error" component of any examinee's observed score can come from many sources: having a bad day, momentary distraction, a lapse of memory or concentration, a lack of motivation, or an experience that inspires unusually high motivation (such as the student reading a passage about her favorite activity), misreading the question, guessing (can be positive or negative), or other random and unknown causes. The fact that errors are random implies that, on average, they cancel each other out and a student's "true score" emerges.

Consider the following situation under CTT: suppose that for a 20-item mathematics test, Dale knows the correct answer to 17 questions, but incorrectly marks the answer sheet for two of those 17 questions. Then, Dale correctly guesses the answer for the 18th question. Under the CTT true score model, Dale's performance might be represented with this logical sequence of statements (each statement being equal to the previous one):

Dale's Observed Score	**=**	**True Score + Error(s)**
"	**=**	**(Dale's Ability) + (Dale's Errors)**
"	**=**	**(answers known: 17) + (mismarked answers: -2) + (correct guess: 1)**
"	**=**	**(17) + (-2 + 1)**
"	**=**	**(17) + (-1)**
"	**=**	**16**

In this example, Dale's observed score is the result of two kinds of non-systematic errors: a mismarked answer sheet and a correct guess. The true score model assumes that *some amount* of error is reflected in the observed score; thus, the number of items answered correctly does not provide the "true" score. True scores, by definition, cannot be directly observed.

To further illustrate, suppose Dale took the same test again and paid closer attention to the marks on the answer sheet, so the second test was free of "marking error." Dale still had the ability to answer 17 of 20 items correctly. Dale also guessed correctly for a question he missed the first time. As a result, the observed score for the second test was different (18 of 20 items correct)—but it still reflected some amount of error.

In theory, if Dale continued to take the same test repeatedly an infinite number of times, we could take the average of *all the observed scores Dale could obtain* and use them to estimate Dale's ability—his true score—for this test (Crocker & Algina, p. 109). In practice, examinees cannot be tested repeatedly, so a test taker's true score cannot be observed; it can only be modeled using a test theory (Kolen & Brennan, 2004, p. 9).

The true score model and its assumptions led directly to the calculation of the **reliability** of the test, a criterion that was all-important in CTT. This vital focus on test reliability then led to the examination of the statistical properties of test items that could *enhance* a test's reliability. Three statistical properties or item characteristics were quickly identified: 1) item difficulty, the proportion of people answering an item correctly; 2) item discrimination, the difference in the item difficulty for a high-achieving subsample of test takers compared to a low-achieving subsample; and 3) item distractor analysis, the analysis of the proportion of test takers selecting each incorrect response option on a multiple choice item.

Researchers quickly learned that the most **reliable** tests were composed mostly of items within a range of difficulty (between roughly .40 and .65) and having particular discrimination values (.3 or .4 or higher) along with **distractors** that were selected by

some reasonable percentage of students. (Item difficulty and discrimination will be discussed in greater detail in section 2-B.)

Advantages/Disadvantages

Classical statistics carry many advantages. The mathematical processes that generate them are relatively straightforward and well understood. CTT procedures have been widely practiced in the field of educational measurement for decades and CTT was once the *de facto* standard scoring paradigm for test and examinee analysis. CTT is still widely practiced and very useful, particularly in test development and item analysis. CTT also allows for test development using smaller sample sizes as compared to IRT, which is often a significant benefit.

It is also important to evaluate CTT in the context from which it emerged. Classical test theory was developed to support norm-referenced interpretation of tests at a time when virtually all educational testing was designed to stratify students and arrange their scores to reflect their relative levels of attainment. CTT was effective in supporting and guiding the development and use of tests for this purpose.

CTT carries the primary disadvantage of creating an inseparable interdependence between *item characteristics* and *test taker characteristics*, both of which are dependent on the sample of test questions and test takers involved. For example, classical statistics may show the overall score level a student demonstrates, but that score is defined only in terms of that particular test. Additionally, item-level statistics can only be interpreted in the context of a particular group of test takers.

Item and test taker interdependencies affect test development in important ways. For example, test construction can become very challenging if examinees taking a test at some future point are significantly dissimilar from the original group of test takers. Measurement experts are quick to note the inseparability of tests and test takers when considering the limitations of using CTT in more modern contexts:

> **Perhaps the most important shortcoming [of CTT] is that examinee characteristics and test characteristics cannot be separated: each can be interpreted only in the context of the other . . . [an] examinee's ability is defined only in terms of a particular test. . . . When the test is "hard," the examinee will appear to have low ability; when the test is "easy," the examinee will appear to have higher ability. What do we mean by "hard" and "easy" tests? The *difficulty of a test item* is defined as "the proportion of examinees *in a group of interest* who answer the item correctly." Whether an item is hard or easy depends on the ability of the examinees being measured, and the ability of the examinees depends on whether the test items are hard or easy!**
> (Hambleton, Swaminathan, & Rogers, 1991, pp. 2–3)

A simple example illustrates the point made by Hambleton et al. above: consider a test in which the average score represents students correctly answering 93 percent of the items. There are two interpretations of this result: 1.) this was a very easy test and perhaps standards need to be raised, or 2.) these students were very capable and perhaps they and their teachers should be acknowledged.

The group-level dependencies associated with CTT also affect test developers in significant ways, because test construction can be very challenging if examinees who will take the test at some point in the future are dissimilar from the original group of interest:

> **[It] is very difficult to compare examinees who take different tests and very difficult to compare items whose characteristics are obtained using different groups of examinees. (This is not to say that such comparisons are impossible: Measurement specialists have devised procedures to deal with these problems in practice, but the conceptual problem remains.) (Hambleton et al., p. 3)**

Other practical concerns or limitations associated with classical test theory include

- the logistics of field testing new items (e.g., groups taking tests in the spring perform differently from groups who take the tests during the fall)
- the logistics of adding questions to item banks (e.g., item data will always be tied to the group to which they were administered)
- different scales are produced when multiple versions of the test are administered with no simple functional relationship between them
- the practical difficulty in constructing tests that are truly parallel (i.e., "a standard that is very hard—if not impossible—to satisfy" [Hambleton et al., p. 4])

A final very important shortcoming of CTT is that student characteristics and item characteristics are reported on *different* scales and are therefore disconnected (quite paradoxically, given their statistical interdependence.) Specifically, student performance is reported in terms of raw scores or some re-expression of raw scores, but item characteristics are reported in terms of percentages or correlations. Knowing that a student made a score of 47 out of 60 possible points does not indicate how that student will do on any given item, nor do we know how a student will do on an item that has a specific difficulty value (such as .61) and a specific discrimination value (such as .33). As we will see in the next section, IRT provides a mechanism for placing people and items on a *common* scale.

CTT is a test theory oriented toward *raw scores* and *fixed test forms*. As we will see in the next section, IRT gives a level of analysis to tests comprised of *different combinations of items*, thus providing practitioners with greater flexibility in terms of test construction and security. IRT allows for the construction of multiple test forms, the rotation of test items to different students, and more secure item banks. While IRT applications have many advantages, classical test theory procedures in combination with traditional equating methods can solve many linking and equating problems and employ statistics more familiar to many measurement practitioners.

2-B. Fundamental Concepts of Item Response Theory (IRT)

Item Response Theory (IRT) refers to a large collection of technical procedures for analyzing test items and **scaling** students based on their item responses. IRT takes into account characteristics of the test items students take and their responses to the items to estimate ability for students. IRT ability estimates take into account students' raw scores, but they also reflect certain characteristics of the test items students answer correctly. For example, using an IRT approach, a student who correctly answers 8 out of 20 items on a difficult test could have a higher ability estimate than another student who correctly answers 12 out of 20 items on an easy test.

Comparisons to CTT are useful for thinking about the basic assumptions of IRT. In the prior section, the *sample-dependent* nature of CTT was highlighted as being a serious shortcoming. By contrast, IRT is a theory of measurement with item-level statistics that are *not* group-dependent. Most importantly, in IRT analyses, people and items can be placed on the same scale and, paradoxically, item characteristics and person characteristics are independent. IRT is sometimes referred to as "Modern Test Theory" to differentiate it from CTT.

Under CTT, raw scores—the sum of all points received for all correct responses on the test—are the basis for determining test taker ability. IRT makes use of total test scores as well, but also takes into account the characteristics of the particular set of individual items on a test form. Thus, IRT looks beyond the score that a student earns and explicitly takes into consideration the characteristics of items comprising the test form.

Basic IRT Models

All IRT models make assumptions about the way test takers will perform on test items. They also assume that to answer correctly, a test taker's response will be based primarily on the test taker's general ability with regard to the subject matter being tested and up to three characteristics or **parameters** of the item, depending on the IRT model used (see Downing & Haladyna, 2006, p. 428). Each of these models can have a different impact on the equating process and key equating decisions.

The most common IRT models in use today are

- the 1-Parameter Logistic Model (sometimes denoted as "1PL" or "the Rasch Model")
- the 2-Parameter Logistic Model (sometimes denoted as "2PL")
- the 3-Parameter Logistic Model (sometimes denoted as "3PL")

The 1-parameter model will be used to illustrate some of the basic features of IRT models, because it provides a simple view of certain major features of IRT models. The 1-parameter model is also frequently referred to as the Rasch model, reflecting the work of Georg Rasch (1980) who first described the fundamental model.

The 1-parameter model is so named because it characterizes a test item in terms of only one feature: the item difficulty. Using this model, an ability is estimated for each possible raw score between 0 and the perfect (100 percent correct) score. All students who make the same raw score are considered to have the same ability, regardless of *which* items they answered correctly to obtain that raw score (Ryan et al., 2002). In contrast, the more general 2- and 3-parameter models characterize items in terms of *difficulty* and *discrimination* (2 parameters) and *difficulty*, *discrimination*, and *guessing* (3 parameters). These terms will be described more fully in the following discussion of IRT's basic concepts.

- IRT is now used in most large-scale assessment programs.
- IRT models apply to items that use **dichotomous scoring** (loosely interpreted as "having either a completely right or completely wrong answer") as well as items scored in ways that allow for partial credit, such as open-ended constructed response items.
- IRT is used in addition to procedures from CTT.

Item Characteristic Curves in IRT, Starting with a 1-Parameter Model

A central concept in all IRT models is the item characteristic curve (ICC), sometimes described as "the basic building block of IRT" (Baker, 2001, p. 7). Two hypothetical item characteristic curves, reflecting the 1-parameter model, are illustrated in Figure 2.1:

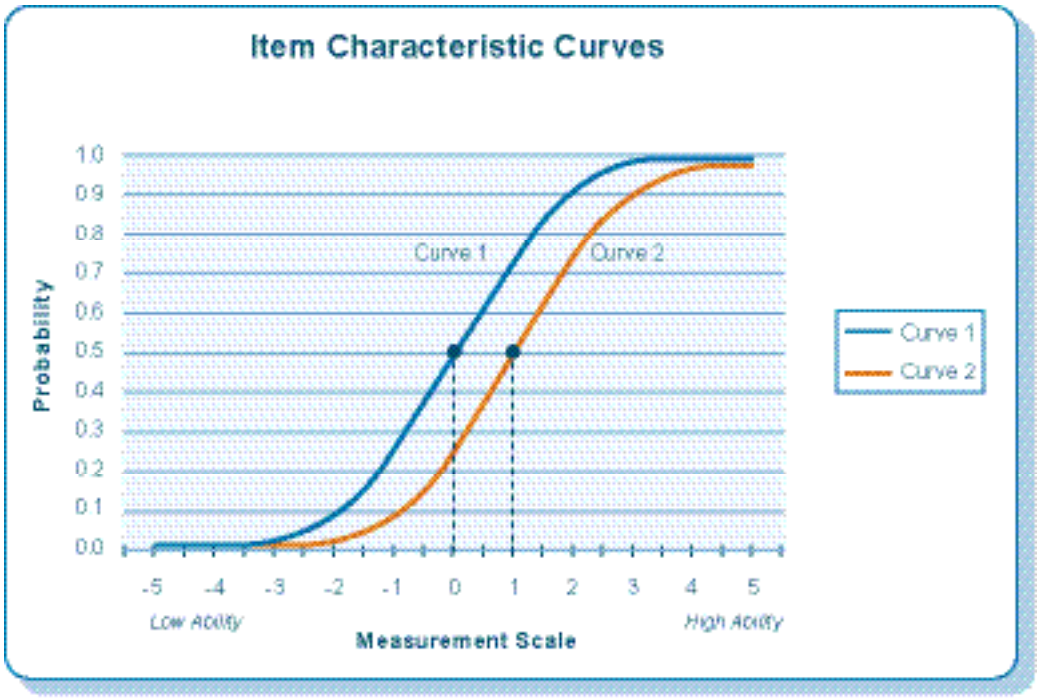


Figure 2.1: Two item characteristic curves (ICCs)

In Figure 2.1, the vertical axis shows the probability of a correct answer, which like all probabilities goes from 0 to 1; the horizontal axis shows test taker ability represented as a number along the measurement scale (which will be explained below); the curves show that students' probability of a correct response goes up as test-taker ability increases.

To find the difficulty of one of the items in Figure 2.1, locate the point along the curve where probability is equal to .5 on the vertical axis, then go down to the horizontal axis to find where that point is placed on the measurement scale. For this illustration, then, the difficulty of the first item represented by Curve 1 is 0. The difficulty of the second item, Curve 2, is 1, showing that it is a harder item; in other words, a test taker would need an ability of 1 on the measurement scale in order to have a .5 probability of answering the second item correctly.

The item characteristic curve is used in IRT to describe the relationship between: a.) the probability of a correct response to the item; and b.) test taker ability. The curve shows the hypothesized and reasonable relationship: examinees with greater ability have a higher probability of answering the item correctly, and those with lower ability are less likely to get the right answer. In Curve 1 of Figure 2.1, students with an ability of -2 are expected to get it right only about 10 percent of the time; students with an ability of 3, about 97 percent of the time.

Item discrimination is the *steepness* of the curve when the probability of a correct answer is .5. An item has only one discrimination, which occurs at .5. Also note that in the simplified examples shown in Figure 2.1, the probability of a correct

response approaches zero as student ability gets lower and lower; in many cases, however, students with low ability may answer an item correctly by chance or by using partial information.

The technical considerations and exact processes involved in obtaining item difficulty and ability values used to construct ICCs are not covered in this handbook. However, it is important to note that under CTT items can be projected or regressed onto the scale used to report students' performance but this is not exactly the same as items and people being inherently located on a single common scale.

The key concept to remember is that under IRT a test taker's ability is based on the characteristics of the particular set of items the student takes and not simply on how many items were answered correctly.

The kinds of information psychometricians use to generate ICCs and the item-level information they represent depends on which IRT model is used to create them. These will be further explained later in this chapter.

Conceptualizing IRT under the 1-Parameter or "Rasch" Model

Once test data has been gathered, there are a variety of procedures that can be used to score student responses in such a way that every student and every item can be placed on a single measurement scale. The technical details of these procedures can be found elsewhere, but the basic concept is important to the fundamental understanding of IRT. This concept of **unidimensional scaling** is illustrated in Figure 2.2 below.

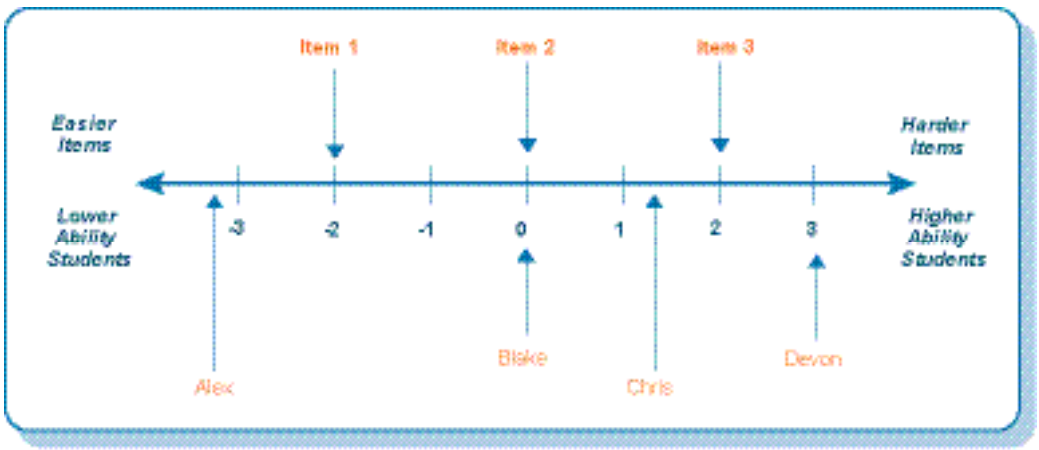


Figure 2.2: Illustration of items and students on an IRT scale

In this illustration, Item 1 is easy, Item 2 is above average, and Item 3 is a difficult item. Alex has low ability, Blake is about average, and Chris and Devon have moderate and high abilities, respectively. Under the assumption of **unidimensionality** we can infer that students like Devon (higher ability) will probably get Item 3 right more often than students like Blake, who will probably get the item wrong. Also, we might say that even though Alex, Blake, and Chris are likely to get Item 3 wrong, Alex is the *most likely* of the three to get the wrong answer. Lastly, we can't support any conclusion about Blake's ability to answer Item 2 correctly because they both occupy the same "average" position on the scale. As a result, Blake's probability of a correct answer is .5—a 50-50 chance of getting it right.

Note, however, the conclusions here are *probabilistic*; they attempt to describe what is more or less probable in a given situation rather than statements of absolute certainty.

The unidimensional scale presented in Figure 2.2, and the probabilistic statements drawn from it, capture the essential elements of the 1-parameter logistical model of IRT. In this model, only a single item characteristic (or parameter) influences the probability that the student will answer correctly.

Figure 2.3 illustrates how an item of average difficulty (such as Item 2 in Figure 2.2) might look when represented in a table of values calculated for these particular students and items. Although the formula for doing this kind of calculation can be complex, key points of interest for Item 2 can be highlighted and compared using Figures 2.2 and 2.3:

Parameter: Item Difficulty

Item 2

Difficulty = 0

Student	Student Ability	Probablity of Correct Answer
Devon	3	0.95
	2.5	0.92
	2	0.88
Chris	1.5	0.82
	1	0.73
	.05	0.62
Blake	0	0.5
	-0.5	0.38
	-1	0.27
	-1.5	0.18
	-2	0.12
	-2.5	0.08
Alex	-3	0.05

Figure 2.3: Table showing student ability for Item 2

Figure 2.3 shows probabilities of correctly answering the question for students of various levels of ability: Alex, Blake, Chris, and Devon. Note that in this table, Blake's probability of answering correctly is .5, as discussed earlier. Blake's ability for this item, then, is placed at 0—just as it was shown in the Figure 2.2.

Figures 2.2 and 2.3 provide a conceptual basis for understanding how these items and students might look when displayed as ICC graphics. If we were to plot the data from this table graphically, we would see the familiar "s-curve," technically known as a logistic curve.

The item characteristic curve for Item 2, as shown in Figures 2.2 and 2.3, could then be represented as shown in Figure 2.4:

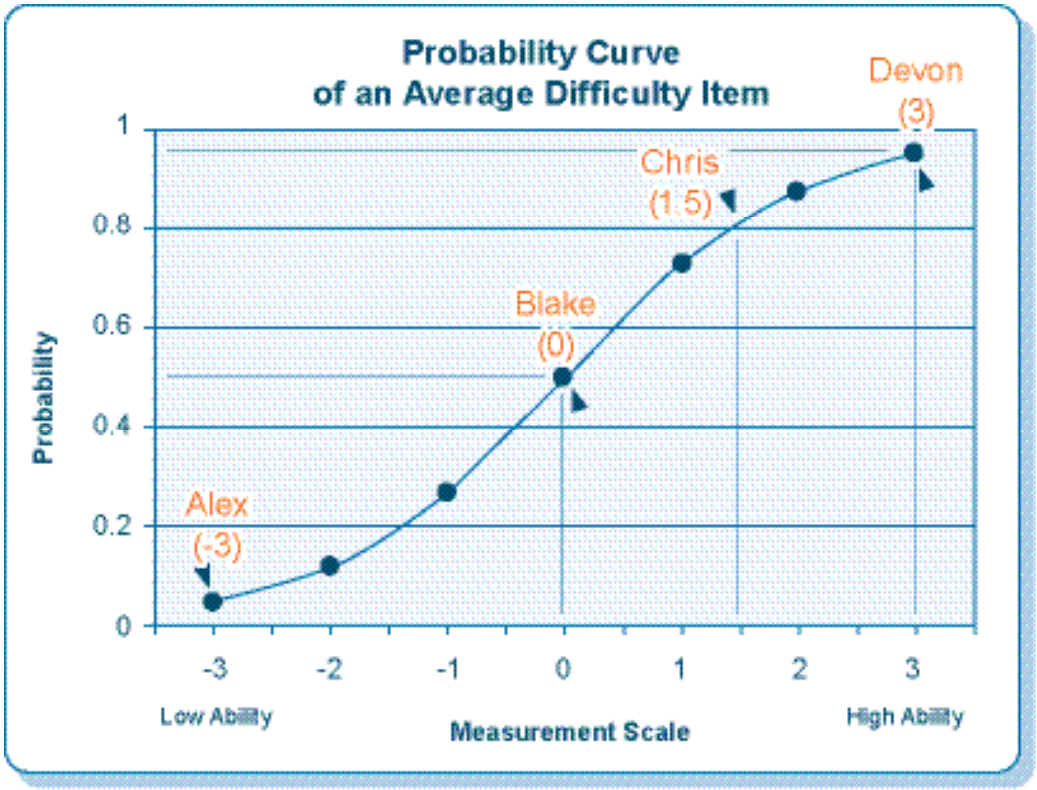


Figure 2.4: Probability Curve for Item 2

In Figure 2.4 we see that the lower the ability of the students on the measurement scale, the less likely they are to answer the items correctly, while students with higher ability are more likely to answer the same items correctly. In this example, Devon has a much higher probability of correctly answering the average-difficulty item (Item 2) than Alex, and Blake has a probability of .5 of answering Item 2 correctly.

Note that the difficulty of an item is always the point on the scale where students have a .5 chance of answering the item correctly. In this example, the average item difficulty is set to 0, but the basis or origin of the scale could be centered at any value (Ryan et al., 2003).

The 2- and 3-Parameter IRT Models

The discussion of IRT thus far has been limited to illustrating items, students, and data that can be described in a simple fashion. The illustrations used capture the basic concepts of the 1-parameter IRT model. Real data, however, does not always follow the 1-parameter model, but once an understanding of the core concepts of this model are in place, readers can gain a better understanding of the more general 2- and 3-parameter IRT models.

The **2-parameter model** uses the item difficulty parameter *and* the item discrimination parameter. In addition to item difficulty, this model includes item-level information that reflects the data showing that some items discriminate more sharply between higher and lower ability students than others. Figure 2.1 shows item characteristic curves for items with identical discrimination; if the ICC was very steep when the probability of a correct answer is .5, the item would be more discriminating. If the ICC is very flat when the probability is .5, the item would be less discriminating. In other words, item discrimination shows how rapidly the probability of a correct response goes up as ability increases.

The **3-parameter model** uses a third parameter in addition to the item difficulty and item discrimination to adjust the lower end of the ICC for possible guessing. By contrast, the ICCs used in all previous examples have the lower ends of the curves approaching a probability of zero, which assumes that no students answer test items correctly as a result of guessing.

An illustration various item parameters, as represented by three separate items with individual ICCs, is shown in Figure 2.5:

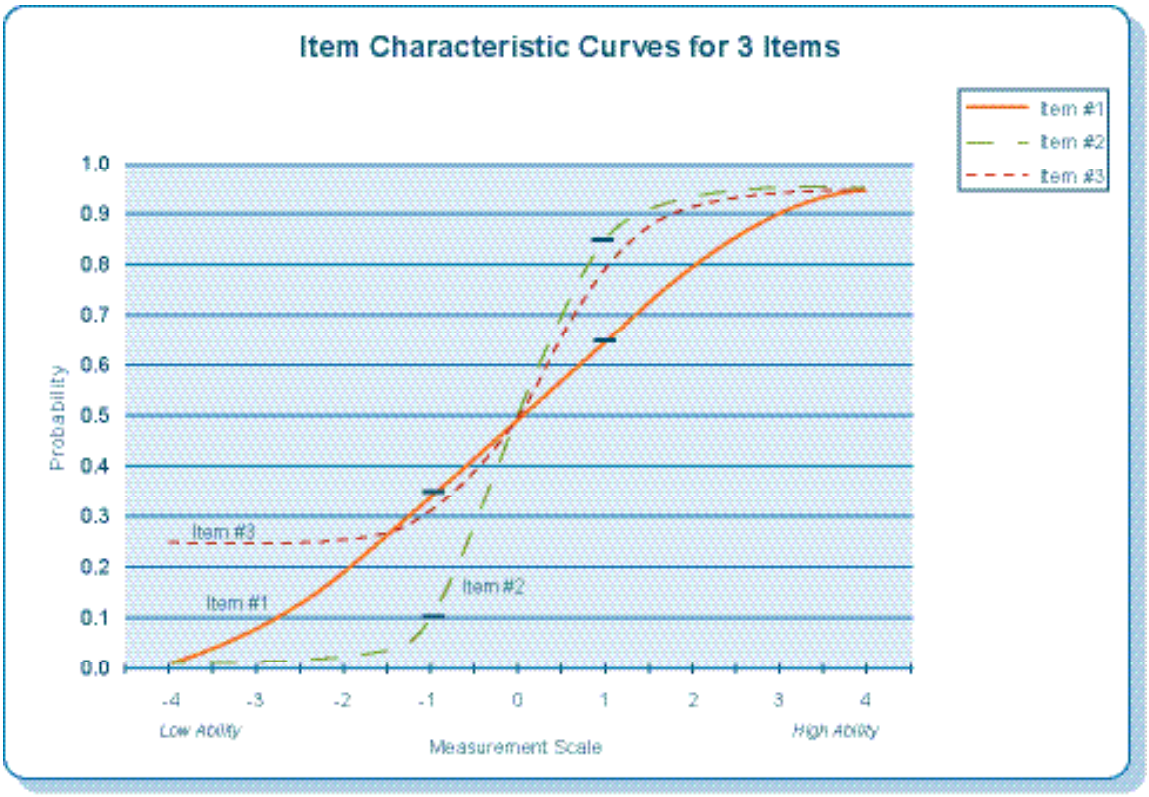


Figure 2.5: Illustration of three item characteristic curves (ICCs)

In the example shown in Figure 2.5, consider students at the ability level of 0: each has a .5 probability of answering each of these items correctly. The difficulty of all three items is the same: each item probability of .5 maps to the same ability level (in this example, 0.)

However, the items vary in discrimination. The ICC slope of Item 2 goes up more sharply than Item 1, which means Item 2 is more highly discriminating. For Item 2, each student at the ability of -1 has about a .1 probability of answering the item correctly, but students at the ability of +1 have a .85 probability of answering correctly (a change in probability of .75). In contrast, consider Item 1: students at the ability of -1 have a .35 probability of answering correctly, and students at the ability of +1 will answer correctly with .65 probability (a change in probability of .30). This difference in the rate of change in the probability of answering correctly is the item's discrimination.

Lastly, the ICCs for Item 1 and Item 2 approach the probability of 0.0 as ability gets lower (moving toward the left-most side of the horizontal scale). The ICC for Item 3, however, appears to flatten out at a probability of about .25. This is the same probability that is expected if students were randomly guessing on items with four answer choices. The 3-parameter IRT model includes the pseudo-guessing parameter to account for this kind of item data, which affects the lower end of the ICCs, although the c-parameter is generally below the theoretical .25 value.

Q: Can we really tell if students are guessing?

A: After students take tests and data is generated for each item and examinee, we are simply left with *data*. By itself, data cannot say with certainty whether a student answering a multiple-choice item has guessed or not. (The only way to really know would be to poll students directly about their answers, and even then we may not get the true account!) Also, if the distractors for an item are reasonably well designed, the actual proportion of low-ability students answering correctly is very often below the probability of chance because students are drawn to attractive but incorrect responses. Therefore, psychometricians sometimes refer to the third parameter of the 3PL model as the "pseudo-guessing" or "pseudo-chance-level" parameter.

Scoring Method

When tests are scored and prepared for analysis with IRT, two kinds of scoring may be employed: number-correct scoring or pattern scoring. The distinction between these two approaches is important to understand before taking up IRT equating.

Number-correct scoring makes no distinction between items that an examinee answers correctly: if Chris answers any 15 of 20 items correctly on a math computation test, and Pat answers some other set of 15 items correctly *on the same test*, both students get the same number-correct score and, most importantly, the same IRT ability. By contrast, pattern scoring takes into account which items a student answers correctly, and the IRT item discrimination and the pseudo-guessing parameter make a difference. Chris and Pat might answer the same *number* of items correctly, but if the items Chris answered correctly tended to have higher discrimination values and lower guessing parameter values compared to the items Pat answered correctly, Chris would have a higher weighted score *and* a higher IRT ability.

Although pattern scoring provides more information about student abilities, various practical problems limit its usefulness for equating, and as a result tests are often scored number-correct even when they are equated using 2- and 3-parameter IRT models (Kolen & Brennan, p. 175). When number-correct scores are used, the equating process under IRT requires another procedure: equating via true scores or equating via observed scores.

Pattern scoring is not used for the 1-parameter IRT model, as all discrimination values are treated as equivalent and guessing is not modeled.

Parameter Invariance and Scale Indeterminacy with IRT Models

An important feature of IRT procedures is a characteristic referred to as *parameter invariance*. The item parameter invariance property of IRT models is essential to all aspects of equating using these models. The invariance property allows a scale to be defined with a fixed origin and with fixed item values. With these values known, subsequent assessments can be linked or equated through items located on the fixed scale. In regard to measuring people, this means that once the scale is fixed, student ability parameters are invariant regardless of which sample of items is used.

The fact that the IRT analyses require values for the parameters to be given some initial fixed starting point is called *scale indeterminacy* (Hambleton, Swaminathan, & Rogers, 1991). The fixed values for IRT scaling can be just about any place along the scale that might be useful. In applications using the 1-parameter model, it is common to center the scale at the average item difficulty. In applications using the 3-paramter model, the scale is commonly set at the mean ability of the students. The scale could be fixed, however, to some other convenient point on the scale, such as a point that defines a "proficient" performance level. Fixing the scale resolves the issues of scale indeterminacy and simultaneously defines a scale on which stable or invariant estimates can be derived.

The stability of different item parameter estimates used to define the scale may be subject to variation due to sampling fluctuations. Item difficulty estimates may be quite stable if the assumption of unidimensionality is strictly met and the data fit the 1-parameter model. Item discrimination and pseudo-guessing may also show a certain degree of stability if samples used to derive estimates are quite similar, but they are subject to variation due to the selection of different samples.

It must be noted, however, that all item and person parameters are estimates with a certain degree of error and the magnitude of these estimation errors also reflects sample characteristics. Large samples of students taking items which have difficulty values well matched to the achievement level of the students have smaller estimation errors than smaller samples—especially if the items are not well targeted to the students.

Parameter invariance is a *theoretical* property of IRT models, which more or less holds true in real settings. Still, variations in item parameter values occur and may be caused by many factors. For example, item parameter values can vary based on small changes in item wording or format. Or there may have been changes in the item's location, changes made to the sequence of items that precede the item, or more focused and targeted instruction. These factors—and many others that are less easily identified—can account for some variation. Tests of fit for IRT models provide some indication of whether the invariance property is likely to be operating with a given data set.

We must also recognize IRT parameter invariance as a property related to samples from the same population, not samples from different populations. We can define the population of "all sixth graders," but this does not mean that all subpopulations of sixth grade students (e.g., ethnic groups, language groups, gender groups) would yield invariant item parameters across the subpopulations.

Numbers, Scales, and Scaling

A critical issue in both classical test theory and item response theory is the question of what numbers or what scale to use in reporting test results. The most intuitive scale, the scale with which all are familiar, is the raw score scale. This is simply the number of points earned (e.g., 23 out of 30, 47 out of 50, etc.). Nearly as familiar is the practice of rescaling raw scores into percentages. With percentage rescaling, 23 out of 30 becomes 76.6 percent, 47 out of 50 becomes 94 percent, and so on.

Both the raw score scale and the percentage scale have several limitations. Most importantly, they are dependent on the particular set of items that make up the test and tend to invite comparisons that are inappropriate and invalid. For example, consider a student with a score on Test A of 45 out of 80, or 56percent, compared to a score on Test B of 71 out of 80, or 89percent. Clearly the student received more points on Test B and had a higher percentage of points earned on that test. However, these results might reflect the fact that Test A is composed of very difficult items (perhaps in the range of Item 3 shown in Figure 2.2) while Test B is made of relatively easier items (in the range of Item 1 in Figure 2.2). Making inferences about students'

knowledge or abilities based on raw scores or simple rescaling of raw scores can lead to incorrect conclusions.

Classical item and test analysis combined with traditional equating procedures can address the shortcomings described here. This approach is the basis for many successful long-standing assessment programs and is well described in various texts (see Crocker & Algina, 1986, and Kolen & Brennan, 2004).

IRT analyses and IRT scales like the one illustrated conceptually in Figure 2.2 provide a useful scale for making inferences about what students know, what they can do, and what makes items more or less difficult. Most IRT software carries out its calculations in a mathematical scale using "logits." A logit scale is mathematically convenient, but it has several shortcomings. First, few people have ever heard of it. Secondly, it has no fixed natural origins (or zero points) for the parameters in the particular IRT model used (1, 2, or 3 parameters). Some initial fixed parameter values are needed to fix this indeterminacy to obtain final parameter estimates from the data.

Once the IRT indeterminacy is resolved by fixing scale values, logit values for item parameters and people can be estimated. These results, however, will still be on a logit scale, which is generally unfamiliar to most educators. As a final step, the logit scale is transformed into whatever reporting scale is desired, typically by selecting the desired scale's mean standard deviation and/or range (see Crocker & Algina, 1986, for a review of procedures for transforming scores).

Common IRT Uses and Applications

IRT-based calculations provide detailed, item-level information about test items that can be useful in selecting items for new test forms. As noted previously, IRT allows test developers to analyze and interpret student- and item-level characteristics and is not limited to test- or group-level analysis. This helps developers to infer a great deal more about how test takers will respond to an item than classical test theory would allow.

IRT is used to

- provide a measurement scale on which both the people and the items can be located
- locate test takers on the same measurement scale even if they have not taken the exact same test forms
- facilitate the careful review of item quality and of the validity of student responses

The most common applications of IRT include

- evaluating and reviewing items and tests
- linking or equating test forms (the focus of this handbook)
- constructing item banks
- constructing equivalent forms from item banks
- setting content-referenced performance level standards
- providing content-referenced score interpretation
- investigating items for differential item functioning (i.e., bias based upon student demographic characteristics)
- supporting Computerized Adaptive Testing (covered in Chapter 5 of this handbook)

IRT provides considerable flexibility in terms of

- constructing alternate tests forms
- administering tests that are well matched or adapted to students' ability level (so that relatively lower-ability students are not overwhelmed while relatively higher-ability students are bored)
- building sets of connected tests that span a wide range (perhaps two or more grades)
- inserting or embedding new items into existing test forms for field testing purposes so new items can be placed on the measurement scale (and eventually be used to construct new test forms)

In general, all of the IRT models provide very similar results in terms of item difficulty and the measurement of students.

The 1PL model is

- logistically simpler to apply, easy to use, and requires smaller sample sizes; however, it provides less precision of measurement
- established, well known, and used in many state assessment programs

The 3PL model

- provides more precision in measuring students by using information about item discrimination and pseudo-guessing
- is established, well known, used in many state assessment programs, and has been used among commercial test publishers for many years
- requires more technical expertise to apply
- requires larger sample sizes relative to other models

It is also important to note that most trained psychometricians would consider the previous illustrations and their explanations to be highly conceptual and this CTT/IRT primer only begins to scratch the surface of the day-to-day use of CTT and IRT concepts and procedures. These may include more advanced models that accommodate partial credit and gridded response assessments analyzed using **polytomous IRT models**. Readers seeking to understand classical and modern measurement theories on more fundamental and technical levels are highly encouraged to refer to the chapter references for further reading.

Chapter Glossary:

Classical Test Theory (CTT)
A body of knowledge that emerged from statistical measurement approaches focusing on observed raw scores; used since the early 1900s.

Dichotomous Scoring
A scoring scheme for which the only possible values of the item score are 0 or 1. Examples of dichotomously scored items are multiple-choice questions that give full credit for one correct answer but no credit for selecting any distractor, true-false questions, and agree/disagree questions.

Distractors
In selected-response questions, the answer choice options which are **not** keyed as the correct answer.

Field Test
A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting and sometimes test form equating. A field test is generally more extensive than a **pilot test.**

Item Parameters
Under IRT, the aspects used in item analysis, including difficulty, discrimination, chance, and person ability.

Item Response Theory (IRT)
One of two key approaches used in modern educational measurement (see Classical Test Theory); sometimes referred to as *modern test theory* or *latent trait theory.* IRT makes use of mathematical models for how examinees with different ability levels will respond to test items with particular characteristics.

Parameters
Particular aspects or ways of looking at items, such as "item difficulty." Technically, it refers to a characteristic of a mathematical model for which statistical values are estimated from data.

Pilot test
An assessment administered to try out new test questions using a representative sample of test takers solely for the purpose of determining the properties of the test or test questions (see Field Test).

Polytomous Scoring
A scoring scheme that allows for partial credit. For example, a short answer test question may allow for a score of 0, 1, 2, or 3.

Reliable/Reliability
The degree to which the scores are dependable, stable, and free of *errors of measurement* (see Chapter 3).

Scaling
The process of associating numbers (or other ordered indicators) with the performance of individual test takers. Raw scores are transformed to *scale scores* using statistical methods. Typically, scales are constructed in ways that will help test users interpret the scores.

Unidimensionality
In non-technical terms, unidimensionality refers to the assertion that a test assesses a single factor or single characteristic of the students taking the test. It is also an attribute of the students in that the unidimensionality claim asserts that students' responses are due to one and only one characteristic of the students. In most assessments, unidimensionality is a matter of degree, not a "yes" or "no" characterization of the assessment.

Unidimensional Scale
In IRT, a single measurement scale constructed by scoring tests in such a way that every student and every test item can be placed on the same scale. The term *unidimensional* refers the basic scale premise of having only one trait to measure. For example, "weight" is a unidimensional trait: there can be more of it or less of it, and it is easily measured by a number scale.

Chapter 3

Basic Terms, Concepts, and Designs for Equating

This section of the handbook defines basic terms and illustrates the fundamental concepts and models for data collection as part of an overall test equating design.

Test linking and equating employs terms and concepts that may be more or less familiar to educators. This chapter begins by briefly describing these key terms and concepts to provide a shared and basic understanding before offering more detailed explanations of common equating designs.

Not all of these terms are used in Chapter 3, but the descriptions below provide an initial exposure to many of the basic ideas and concepts used in equating. Each of these terms will be explained in more detail in this chapter and subsequent chapters.

Anchor Items/Linking Items – The terms anchor items and linking items are often used interchangeably. Anchor/linking items refer to a single set of items that appear on two or more tests forms. These items, common to two or more forms, are said to serve as "anchors" that fix the measurement scale on which the test forms are connected (equated). The items common to two or more forms can also be described as the "links" which connect the forms together onto a common scale.

Appended/Embedded Anchors – Anchor items that appear at the end of a test form are described as appended anchor items. Anchor items that appear at various positions throughout a test form are referred to as embedded anchor items.

Field Testing – The practice of administering test questions that often don't count toward student scores in order to check item quality in general and to obtain working values for IRT item parameter estimates. In the context of equating, field testing is often used to develop an initial equating to yield a set of equated forms or an initial item bank. *The usefulness of the IRT values and equating that result from field testing depends on how similar the context and dynamics of the field testing are to the actual operational testing situation.*

Form-to-Form Equating – Test forms equated through a series of pairwise equatings. For example, Test Form A is equated to Form B through a set of linking items common to Forms A and B; Form B is equated to Form C through a set of linking items common to Forms B and C, Form C is equated to D through items common to Forms C and D, and so forth. In theory, all forms are placed on a common scale through this sequential process.

Horizontal Equating – Refers to the most common need in statewide assessment programs—to have grade-level scales and performance standards remain stable within each grade. Equating test forms within the same grade level or age range is referred to as horizontal equating and is the same as form-to-form equating.

Item Bank – Any set of items might be called an item bank, which may include the actual text of the questions along with any graphics, item attributes, answer keys, etc., along with item parameter values obtained though field testing. Items in the secure item bank are available only to test developers for the purpose of constructing test forms; these item banks are secure and are not to be confused with other sets or banks of items used to build and administer interim, benchmark, or formative assessments at the state or district level. In the context of equating, however, there is added special meaning in that all items in the bank have been placed onto a common scale through some form of linking.

Item Parameter Drift – IRT parameter estimates for item difficulty, discrimination, and the pseudo-guessing parameter are generally used as if they remain stable or constant when items from an item bank or another test form are used on a newly constructed test form. In some cases, however, the IRT values change or drift away from their bank values and any substantial item parameter drift can compromise equating when IRT methods are employed.

Multiple Forms, Common Anchors – A common set of anchor items appears on all forms to be equated. For example, if Forms A, B, C, and D are to be equated, all four forms would have the same set of anchor items.

Pre-equating – A common application of pre-equating involves constructing a new test form from items contained in an item bank. The new form is constructed to match IRT difficulty specifications as well as content and format specifications. Before the new test is given (i.e., pre-administration) a scoring table is constructed based on the existing IRT bank values to show the scale score associated with each possible raw score on the new test.

Post-equating – Tests that are equated after administration may be constructed using existing IRT bank values, but the final equating of the test to the official operational measurement scale is done with live data from the test administration. It is desirable to do post equating on data from the full population, but when reporting schedules require expedited processing, the post-equating is done on a special "early return" sample of schools chosen to match relevant characteristics of the student population.

Spiraling Test Forms – A method of distributing multiple test forms within a student group such as a classroom or school. Spiraling usually occurs when multiple test forms, such as Forms A, B, C, and D, are randomly distributed within the group of test takers. This is often done by packaging the forms in a sequence such as ABCDABCDABCD, etc, and then distributing the forms sequentially.

Vertical Scaling – Refers to an application in which an item bank or a set of test forms is developed and equating procedures are used to create a scale that spans a range of ages or grades. Although this is sometimes referred to as vertical equating, often it does not meet the requirements associated with the strict definition of equating (e.g., context and construct equivalence). However this may be accurately be described as *linking* test forms from one grade level to the next (Patz, p. 6, 2007). In this handbook, vertical linking and vertical scaling will be used synonymously.

3-A. Equating Designs

The practice of equating two test forms is a practical matter that requires actual scores as a starting point. The scores (data) used to perform linking and equating calculations are collected according to established principles known as data collection designs or equating designs. The choice of what scores to use must be very purposeful and deliberate, and it must satisfy certain requirements to be technically defensible.

Before any equations can be employed or any calculations can be made, a sound and appropriate design must be selected to ensure the data gathered is suitable as a basis for equating. For example, most psychometricians would probably agree that linking test forms using only scores from one elementary school in Oregon (Test Form A) and scores from every school in five counties in Florida (Test Form B) is an indefensible data collection design.

Questions we might pose to help clarify the process of creating/selecting an equating design include

- **What opportunities and resources are available for producing field test forms, embedding items for linking/equating on a regular test administration, or conducting a special stand-alone equating study? Which of these are not feasible?**
- **What latitude is there in arranging for the collection of different samples or for spiraling test forms within classrooms, schools, or districts?**
- **What data would or could be available to support the equating process?**
- What are the requirements for the public release of test items? Are items used for equating tests but not for scoring students exempt from these requirements?

Answers to these questions are an appropriate place to begin because practical constraints may compromise the value of even the most soundly conceived equating design.

All equating designs face logistical, financial, educational, and statistical constraints.

The information in the following sections represents an abbreviated survey of information derived from resources that provide deep analyses of several of the most common equating/data collection methods:

- *Equivalent Groups (Random Groups) Design*, which is used in many large scale assessment programs
- *Single Group Design*, which provides the conceptual basis for other designs
- *Single Group Design with Counterbalancing*
- *Anchor Test Design*, which utilizes concepts and practices that are quite common in many statewide testing programs

Readers interested in a more thorough discussion of linking and equating designs and other related topics will find ample reference materials available, such as *Educational Measurement,4th Ed.*, (Brennan (Ed.), 2006), *Test Equating, Scaling, and Linking: Second Edition* (Kolen & Brennan, 2004), and *Linking and Aligning Scores and Scales* (Dorans, Pommerich, & Holland, 2007). These and other sources listed in the chapter references are highly recommended for further reading.

Equivalent Groups (Random Groups) Design

The Equivalent Groups design (also referred to as the Random Groups design) is built on the principle of **random sampling**. If two random **samples** of sufficient size can be obtained from a testing population, these samples can be said to be functionally *equivalent* in terms of student achievement. Or, as stated in *Educational Measurement, 4th Ed.,* the two groups are "as equivalent as two random samples from the same population can be." (Brennan (Ed.), 2006).

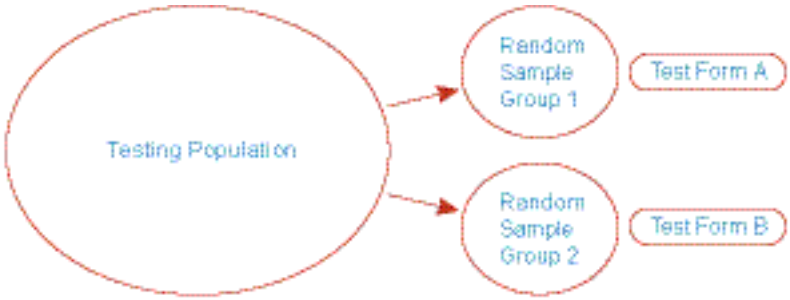


Figure 3.1: Equivalent Groups (Random Groups) Design
Two random sample subgroups are used; each subgroup takes a different test form. The Test Form A and Test Form B groups are said to be randomly equivalent.

This design often utilizes the practice of spiraled test forms to create the random sample groups. In the example in Figure 3.1, Forms A and B might be packaged in A/B/A/B order with instructions directing test administrators to distribute forms A and B alternately from one student to the next. This random assignment of test forms creates two random sample groups within each classroom.

Advantages/Disadvantages

An advantage of this design is its relatively low impact on individual test takers: no single student is required to take *both* Test Form A and Test Form B. This equating design therefore has the advantage of avoiding the problem of **order effects** that are sometimes associated with the other designs. Order effects refer to differences in student performance due to test-taking sequence; the experience of taking Test Form A can potentially alter performance on Test Form B in designs that require students take two tests.

A shortcoming of this approach arises if randomization is achieved by spiraling all forms within classrooms and schools. In these situations, all items from all forms have been exposed in a single setting, and the risk of compromising the entire set of items and forms is increased.

Another potential disadvantage of this design is the need to obtain relatively large sample sizes[1] in order to show that the items in Test Form A and Test Form B are stable and perform reliably. In some cases, this requirement makes the use of this design difficult or impractical.

This data collection design supports linear, equipercentile, and a variety of IRT equating approaches (these will be covered in greater detail in Chapter 4).

Single Group Design

The Single Group design forms the theoretical basis for the Single Group with Counterbalancing design (described on the next page), which is far more common. As such, it is a useful starting point for understanding many of the most common equating designs.

The Single Group design requires that the same test takers take both Test Form A *and* Test Form B. This design could be used if an entire tested population took both test forms to be equated, but in practice it is more feasible to use a randomly-selected subgroup from the testing population.

Figure 3.2: Single Group Design
One random sample subgroup is used;
the subgroup takes both Test A and Test B.

The Single Group design assumes that certain activities that might affect student performance on Test Form A when compared to Test Form B are negligible (such as taking lots of practice questions only for Test Form B or using the results of Test Form A to guide specific areas of study in preparation for Test Form B).

Advantages/Disadvantages

One advantage of the Single Group design is that there is little question about whether or not students' abilities in the sample are similar; they are more than just similar—they are considered to be essentially the same. In technical terms, this is referred to as *controlling for differential examinee proficiency* (Brennan, 2006).

The Single Group design also has practical uses other than typical form-to-form linking or equating. For example, it may be utilized to create shorter versions of longer test forms: after a group of test takers has completed the full-length version of the form, it may be possible to eliminate some items and then link the shorter version to the original.

A disadvantage of the Single Group design is that administering two separate tests to the exact same subgroup of students may not be practical. Few state testing programs can arrange for students to take two complete test forms. This design also exposes *all items* on two test forms to examinees, which may be undesirable due to test security risks, the potential for testing fatigue, and other reasons. The single group design typically requires a special administration that might differ from a standard operational administration.

The potential for order effects is usually too significant to warrant the use of the Single Group design in large-scale statewide testing programs. As a result, equating designs that use this data collection method should *always* be counterbalanced.

Single Group Design with Counterbalancing

When scores from two tests are being equated or linked using a Single Group design, it is important that the order of administration is **counterbalanced**; that is, a randomly selected half of the group should take Test Form A first and the other half should take Test Form B first.

A variant of the Single Group design, the counterbalanced method uses two randomly sampled groups instead of one single group. It may help to think of this design as being "two single groups" as used in the Single Groups design, except that the **counterbalanced** design requires that each group take the tests in a different order:

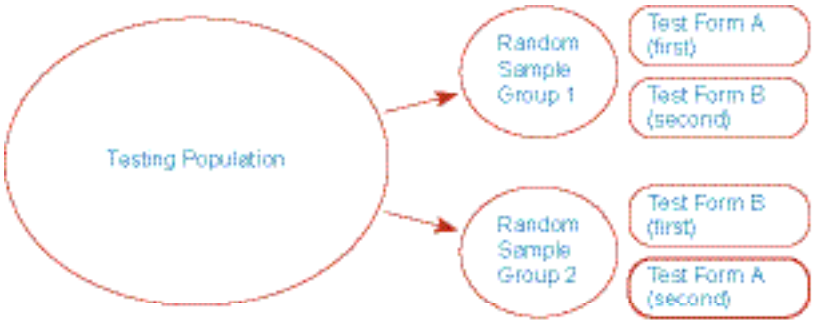


Figure 3.3: Single Group Design with Counterbalancing
Two random sample subgroups are used; each subgroup takes two tests, but in a different order.

Since the same students take both forms of the test, any differences in scores can be attributed to differential difficulty of the tests (assuming that the tests have been constructed to be **parallel** in content).

Advantages/Disadvantages

The advantage of the Single Group counterbalanced design is that it controls for **order effects**, where the experience of taking Test Form A can potentially alter student performance on Test Form B.

However, the primary disadvantage of the Single Group design, mentioned previously, is also present in this design: administering two separate tests to the exact same subgroup of students is impractical.

The Single Group design supports equating procedures such as equipercentile, linear, and IRT equating—all concepts that will be explained in greater detail in Chapter 4.

INFO

Counterbalancing draws its name from the way the alternating order is used to "counter" the order effects of the Single Group design. Psychometricians may recommend or use this design in cases where the impact of order effects is too great to ignore.

Anchor Test Design

The Anchor Test design[2], also referred to as the Common-Item Nonequivalent Groups Design (Kolen & Brennan, 2004) or the Non Equivalent groups with Anchor Test (von Davier, 2004) involves the use of a subset of test items ("anchors") in each of the tests forms to be equated.

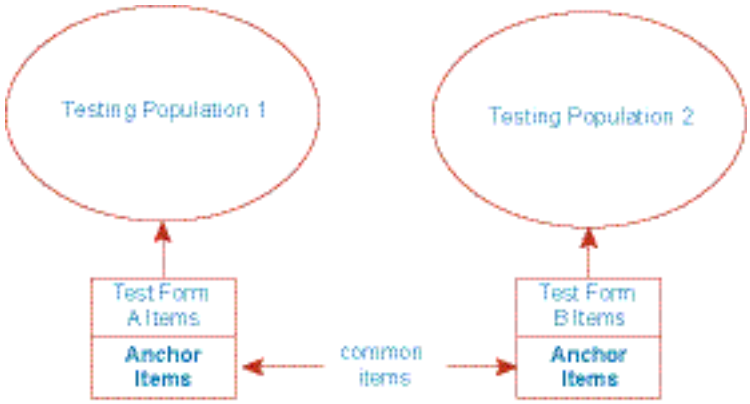


Figure 3.4: Anchor Test Design
Each testing population takes only one test, but each test form shares a common set of test questions (anchor items).

When using the Anchor Test design to equate test forms, the psychometrician's job is to discern whether any differences between the two populations' overall results are due to the students being different, the test items being different, or both. Kolen & Brennan (2004) refer to this task as separating *group differences* from *test differences*. Significant *test* differences might require a closer look at various influences at the item level (e.g., scoring differences, form construction issues or irregularities, significant variations in test administration practices) that could complicate the equating process and the comparability of test scores. Significant *group* differences may require closer scrutiny of the sampling methodology, or an investigation into any underlying factors affecting an entire group (for example, a major natural disaster that affects one of the testing populations).

INFO

When two or more test forms are spiraled within classrooms, the term "nonequivalent groups" does not apply—the act of spiraling creates a sampling of "automatically equivalent groups" by randomly distributing test forms that vary from student to student within the classroom itself. (See **Spiraling** section in this chapter of the handbook for more detail.)

Anchor Test Representation

The selection of items intended to serve the role of anchors is particularly important. The proportional content representation of items in the anchor set should be similar to the proportional content representation of the entire test form, even to the point of considering the anchor set to be a "mini-version" of the full test form (Kolen & Brennan, 2004, p. 19). Although some IRT approaches work reasonably well even if this guideline is not strictly followed (Sinharay & Holland, 2007), in practice the "mini-test" principle is still desirable.

A simplified example of the "mini-test" concept is illustrated in Figure 3.5:

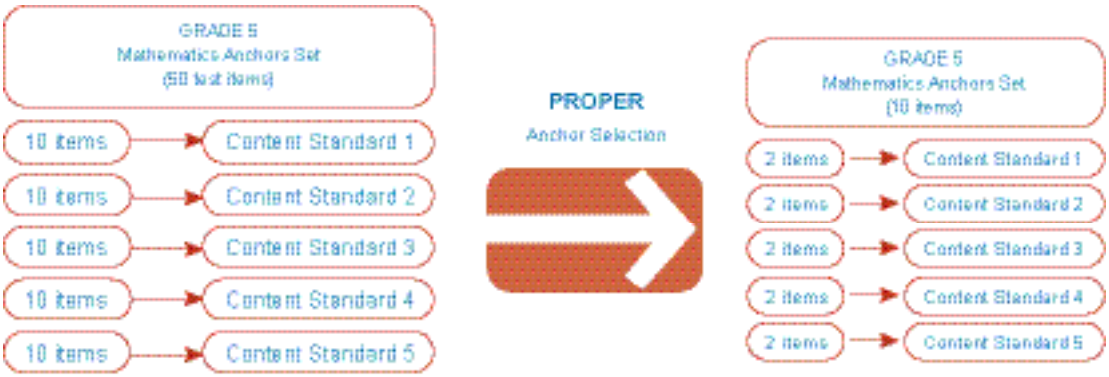


Figure 3.5: Proper selection of a "mini-test" anchor set
An example of a well-selected set of anchor items. Note the similar proportion of items for each content standard when comparing the full test form to the anchor set.

Anchor item selection is typically more complex than the simplified examples shown in the figures 3.5. For example, other considerations are often part of the mix—item difficulty, the coverage of objectives *within* content standards, items using a common stimulus, et cetera. Psychometricians generally expect to see at least 15 to 20 anchor items for longer test forms.

Anchor Item Locations

In practice, anchor items can be located anywhere on a test form, and may or may not contribute to students' scores. The following is a list of common anchor item locations and functions. The first two illustrate *internal* anchors, and the third is an *external* anchor set [3]:

- Anchor items interspersed throughout the test, and contribute to students' score. This is often called an embedded, internal anchor set.
- Anchor items appear together as a block at the end of the test form, and contribute to students' scores. This is sometimes referred to as *appended* since the anchor set appears at the end of the test. In either case, the items are internal anchors because they appear as part of the test form and contribute to students' scores.
- Anchor items appear as a separate form or testing session, and do *not* contribute to students' scores. This is referred to as an appended, external anchor set.

An example of an embedded anchor test model with interspersed items is shown in Figure 3.6:

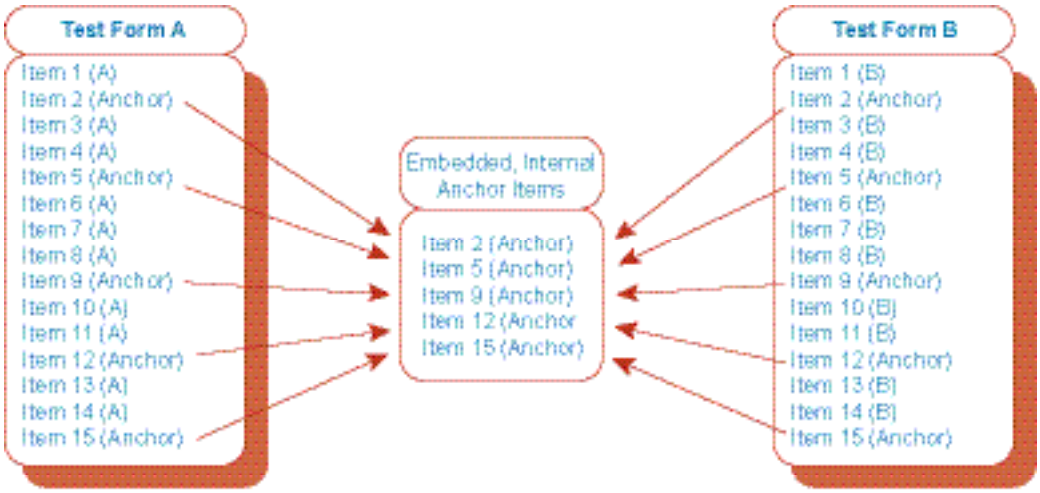


Figure 3.6: Internal, embedded anchor items
An example of a five-item embedded anchor set on two different test forms.

In the example shown in Figure 3.6, each anchor item occupies the exact same item number (location) on both forms. While this is ideal, in actual practice anchor items are often acceptable in *similar* positions on the two tests. For example, if Item 5 in Form A is used as an anchor, it might be positioned as Item 6 in Form B (for a variety of reasons).

Figure 3.7 shows a simplified example of an internal, appended anchor test:

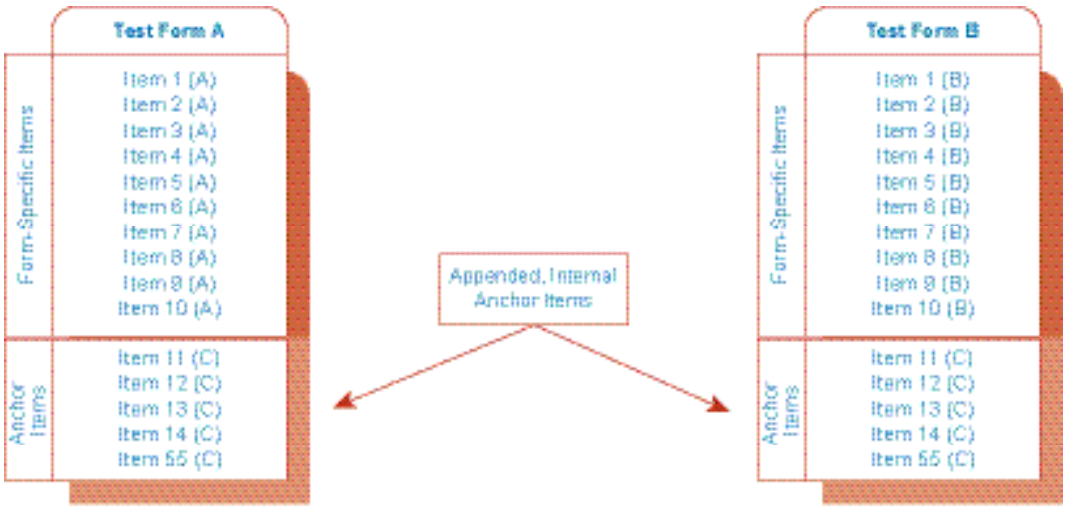


Figure 3.7: Internal, appended anchor items
An example of a five-item appended anchor set ("C") for two test forms. The anchor set is part of the test and contributes to students' scores.

This design for an anchor set is not desirable, since students' performance on the anchor items may be susceptible to fatigue or declining motivation. It has the advantage of students' spending time to answer anchor items which actually contribute to their overall test scores.

Figure 3.8 shows an example of an external anchor test model, with both forms using a separately administered Part 2 as the external anchor set.

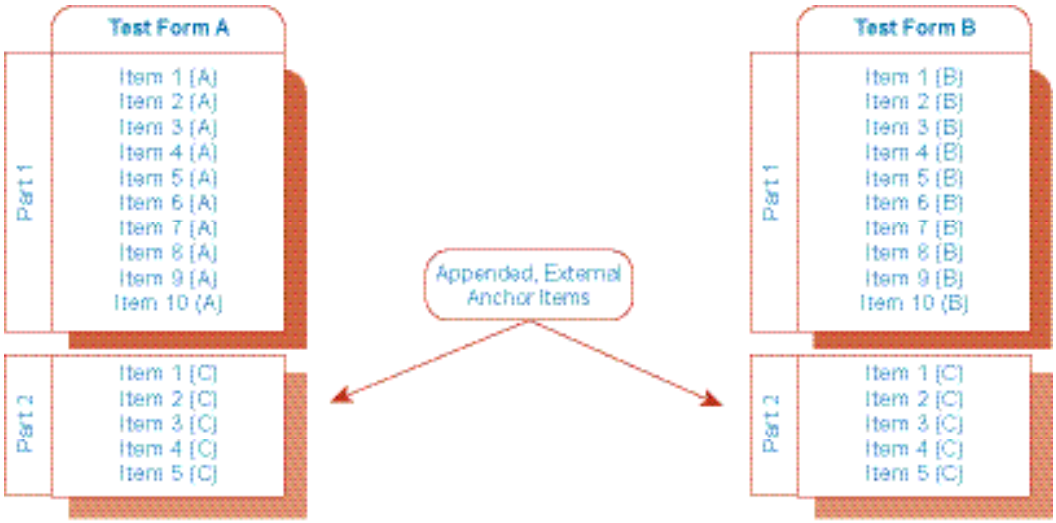


Figure 3.8: External anchor items
An example of a five-item appended anchor set ("C") for two test forms. The appended anchor set is external, does not contribute to students' scores, and is timed separately.

An external anchor set has the advantage of giving students more time to focus on test items that contribute to their scores. The disadvantage is that students may lack motivation if they perceive that these external anchor items "don't count."

Other Advantages/Disadvantages

One advantage of using the Anchor Test design is that **equivalent groups** of test takers are not required in order to establish a basis for linking and equating the two test forms. For example, in Figure 3.8, Test Form A could be given to this year's students and Test Form B could be given to next year's students.

A second advantage of the Anchor Test design is that because it requires only one test administration per year, it can be utilized in conjunction with test schedules that are commonly employed anyway—as opposed to the Single Group design with or without counterbalancing, which would require each test taker to take more than one test form (Kolen & Brennan, 2004).

In some state assessment programs where embedded anchors contribute to the test taker's score, they tend to provide for anchor set content that more closely matches the overall composition of the test form. Also, because the anchor items are often spread throughout the test form in ways that make them impossible for test takers to identify, they are less likely to be skipped (Brennan, 2006).

An advantage of appended anchor items is that they reflect a more modular design that would facilitate the logistics (e.g., printing, test book page production) for an entire group of non-anchor test items to be publicly released with relative ease. For example, Part 1 of Test Form A as shown in Figure 3.8 could be publicly released without compromising the security of Test Form B. Brennan, et al., also note in *Educational Measurement, 4th Ed.,* that this design can lessen the impact of a security breach.

Disadvantages of the Anchor Test design include the statistical analysis requirements of the design and the potential for **context effects** (changes in student performance as a result of other items taken before the anchors). Although the anchor items are

common to both test forms (by definition), the items surrounding the anchors are different for each test form. For example, the non-anchor items on Test Form A may subtly inform students how to perform better on the anchor items. The result may be significant differences in student performance with regard to the anchor items. To control the potential for context effects, Anchor Test designs must be implemented with a careful eye toward specifying clear and unambiguous rules for constructing the test forms and the rules for anchor item placement.

The primary disadvantage of embedded anchors is their potential for context effects and the issue of security breaches (Brennan, 2006). A security breach of a test form with internal anchor items would jeopardize the validity of score results in a way that is difficult to contain since the entire anchor set is present in all forms to be equated. Security breaches can be extremely difficult to control because the physical security of test books is often beyond the scope of those who design, publish, and administer the tests. Alternate test forms that contain very few or none of the same items as the operational test are sometimes available to help lessen the impact of security breaches (these test forms may called **breach forms** for this reason).

In some state assessment programs, anchor items are used to equate forms and replenish item banks but are not used to score students. In these cases, students take items that "don't count" and there is some concern that the unscored anchor items may slow down, fatigue, or discourage students—thus influencing their performance.

In summary, anchor set selection guidelines include the following:

1. *Mini-Test.* Anchor sets should represent a "mini version" of the overall test form.
2. *Similar location.* The anchor items in Test Form A should appear in Test Form B at about the same location (item number).
3. *No alterations.* The anchor items should appear exactly the same in Test Form A as they do in Test Form B and should not be reworded or present answer choices in different orders, different artwork associated with stimulus material, different directions, or any other alteration that might affect student performance from one test to the other.
4. *Item Format.* When possible, the anchor item set should use approximately the same proportional mix of selected response, short answer, and extended response item formats as used on the test form overall.

A careful analysis of anchor item behavior *before* equating is also a requirement for using the Anchor Test design. Also, a scrupulously careful review for any possible item interactions reflects standard practice that should be employed in designing and reviewing any operational test form

3-B. Related Concepts and Procedures: Item Banking, Matrix Sampling, Spiraling

Item Bank Development

Most large-scale assessment programs engage in the process of assembling or constructing a large set of items known as an "item bank" or "item pool." These secure item banks are frequently maintained by testing contractors. In the interest of security, they may also be carefully segregated from any other item banking systems used to produce interim, benchmark, or formative assessments at the state or district level, although in theory items from a secure item bank could be released and used for almost any alternate purpose (e.g., most statewide assessments provide for some form of public release of test content, which could be incorporated into other district- or classroom-level assessments).

Essentially, any of the equating or data collection designs and procedures used in this handbook can be used to establish the initial item bank.

Recall from Chapter 2 that unlike CTT, IRT does not require *item characteristics* and *test taker characteristics* to be dependent upon a particular test sample. Instead, IRT allows psychometricians to estimate test taker and test item characteristics when they take any particular test item or set of test items. This analysis of item-level information distinguishes IRT from CTT and makes efficient and effective item banking possible.

Characteristics of the student samples are a critical aspect to consider in collecting data for a secure item bank. Many IRT procedures are robust with respect to sample characteristics, but it is prudent to establish the bank using samples that are as similar to the intended population as possible for reasons related to both psychometrics and public relations. Using all possible examinees to collect data can provide strong validity evidence. If census data cannot be used, carefully selected samples are strongly recommended. Furthermore, it is useful to construct samples so that the assumption of randomly equivalent groups is justifiable, regardless of the equating procedures to be used. Ideally, test forms used in establishing an item bank should also be randomly administered to the smallest sampling unit possible (e.g., the classroom) to help the samples be as randomly equivalent as they can be; however, psychometricians and other decision makers should also weigh this approach against the risk of exposing all items to a single school or school district.

It is essential that traditionally underrepresented groups be included in the sample used to establish an item bank. When feasible, the generalizability of bank equating procedures should be verified by replicating the bank analyses for subgroups of the population. This may require over-sampling proportionally small groups (e.g., Native American, Pacific Islander, et cetera.)

Refer to Chapter 5 of this handbook for a more detailed look at item banking and bank development procedures.

Matrix Sampling

A pool of test items can be separated into smaller blocks and randomly assigned to different test takers. This use of **matrix sampling** reduces the testing burden on individual students but allows for a relatively large number of items to be administered. Because each test form is linked to at least one other test form, a "chain" can be created to link all forms together. In field testing situations, this practice maximizes the number of items that can be tried out by distributing the items over large groups of test takers.

A basic example of matrix sampling is shown in figure 3.9:

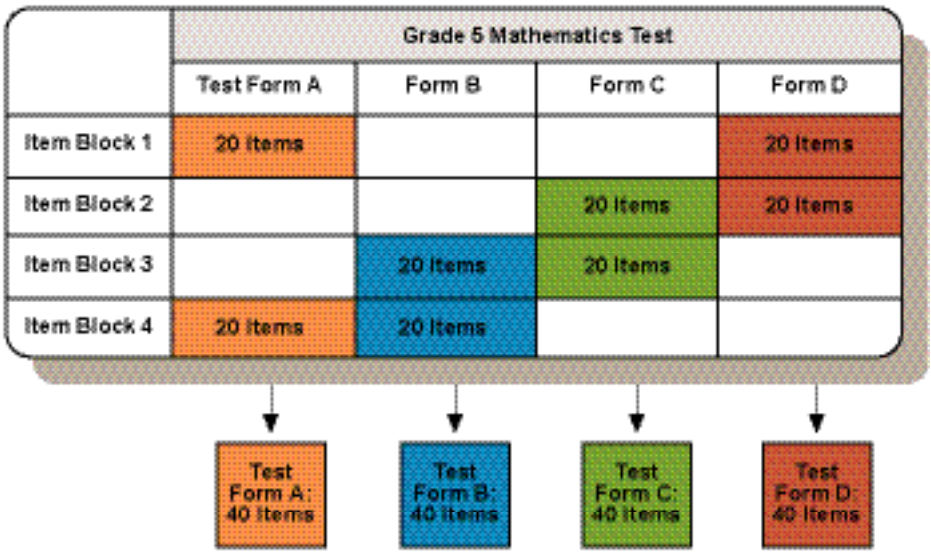


Figure 3.9: Matrix Sampling

Figure 3.9 shows a matrix sampling model in which 80 items are field tested, but no individual student is required to take more than 40 items during his/her test session. In theory, Test Form A could be administered at one school, with Form B administered at another school, and so on. The number of test forms to be created is limited only by the constraints of the final sample size needed for each form.

In practice, however, matrix sampling is typically utilized in conjunction with spiraling methods (see below) within schools and/or classrooms so that two students sitting next to each other may take two entirely separate test forms during the same testing session.

If the total number of students is fixed, then the greater the number of test forms to be matrix-sampled, the smaller the sample size must be for each particular form. This is important because smaller sample sizes can limit the choice of equating methods; if sample sizes are too small, equating cannot be done effectively, if at all (Dorans et al., 2007, p. 67). Therefore, a reasonable estimate of the number of examinees who are likely to take any given test form must be taken into consideration when employing the matrix sampling method.

Spiraling

Spiraling refers to the way test booklets are assembled, packaged, delivered to testing sites, and distributed to students. In a spiraled test administration, test takers are randomly assigned different test forms which are distributed in such a way that each test taker's assigned form is different from the adjacent test taker's form.

Figure 3.10, for example, shows a common-item equating design which uses test forms in a spiraled administration within Pat Jones's small class of 12 students at Jefferson Elementary (isn't Pat lucky?). Each test form contains a set of 15 common items as well as a set of 25 matrix-sampled items that are unique to each form (also see Figure 3.9):

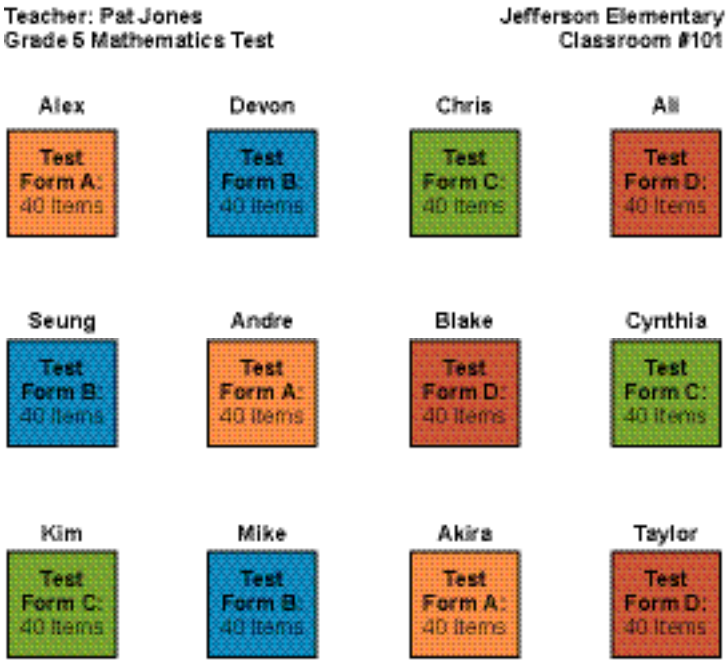


Figure 3.10: Test form spiraling within a classroom

According to Kolen & Brennan, the spiraled method of distribution typically leads to comparable, randomly equivalent groups taking test forms A, B, C, and D (2004, p. 13). In most settings, it is ideal to spiral blocks of items *within a classroom*. The reason for this is because students within classrooms are likely to be more similar (e.g., equivalent) then students across classrooms, schools, or districts. However, if this is not feasible, spiraling within schools or districts are the next preferred options.

Advantage/Disadvantages

A key advantage of spiraling is that programs can field test large numbers of items at one time. And by spiraling at the classroom level rather than at the school or district level, the sample obtained for each set of new items is "as random as it can be."

However, spiraling within the classroom creates challenging logistics. Multiple forms of the test often mean that test administrators must deal with questions that arise from having several different test forms in use during test administration. In addition, great care must be taken in matching the appropriate score key to the correct matrix-sample form.

Some assessment programs utilize the concepts of matrix sampling and spiraling somewhat differently than the conceptual models presented here. For example, the National Assessment of Educational Progress (NAEP) has used a matrix design in which the complete set of items is divided into non-overlapping blocks and then combined into booklets so that each block is matched with another block. This minimizes the testing burden for individual students, but still permits estimates of proficiency at the state level. However, NAEP is not used to produce results for individual students.

The terms matrix sampling and spiraling are related to—but should not be confused with—the larger concept of randomly sampling from a given population. In the example shown in Figure 3.9, the matrix design is intended to randomly assign sets of *items*; however, it is also possible to use a matrix concept to sample student populations. Thus, it can be helpful to think of spiraling and matrix sampling as basic design concepts that are employed in the service of collecting viable data.

3-C. Imprecision in Measurement

Equating procedures are applied in a broader measurement context in which there are several sources of imprecision. In many equating situations, samples of students are used to do special equating studies. Even when using exactly the same procedures to randomly select samples, this process introduces a degree of variation because the two samples can still differ. These differences in the characteristics of randomly selected samples reflect *sampling error*. Additionally, the equating procedures themselves introduce some imprecision, inconsistency, or error in the equating process, and the test form itself has a certain degree of measurement error.

Practitioners should always insist on technical documentation that reports the precision of the various technical procedures used to develop and support their assessment program. Information about technical precision should be used to evaluate various procedures and help in making decisions about students and school programs.

Random Error

Ambiguous or imprecise language often results in misconceptions within the educational assessment community, which comprises a diverse set of professional disciplines. To a psychometrician working to link or equate test forms, the concept of **error** is interpreted through the lens of statistical training—that is, *as a concept or construct* (such as the difference between the observed or expected value and the true value of something) encountered in the process of solving problems. To policymakers and others who may not have statistical training, error is sometimes interpreted as *a result or outcome,* such as "the result of something that has gone wrong" or "the state of being incorrect."

It is important to note that when psychometricians refer to error in the context of equating, they are usually referring to the statistical concept of error. In statistical contexts, error does not mean a mistake has been made or some procedure has been applied incorrectly. Rather, error represents a concept akin to "wandering values," not mistakes. As such, error can be calculated and described by psychometricians as

the level of precision or certainty that test scores, samplings, parameter estimates, or equating procedures are known to produce.

Most people are familiar with the concept of sampling error in the context of public opinion polls, which might report a percentage of people who support a candidate or position with a margin of error ±3 percent.

In general, it is useful to note that increasing the sample size reduces the amount of random error introduced by sampling fluctuations.

Chapter Glossary:

Breach Form
An equivalent (equated) test form created for use when a security breach would otherwise render test results invalid for a student (or group of students). For example, if a test administrator found that one classroom had unrestricted access to one part of a standardized test several days before the exam was administered, a breach form could be used instead. Not all security breaches can be contained via the use of breach forms, however.

Construct
The underlying theoretical concept or characteristic a test is designed to measure.

Context Effects
Within test forms, complications that arise from the variation of questions and formats in different test forms. For example, an item in Test A adjacent to an **anchor item** may help test takers answer the anchor item correctly, whereas students who take Test B may have a different adjacent item that has no effect on anchor item performance. In a more general sense, context effects may include any uncontrolled and sometimes unknown factors that influence students' behavior.

Errors of Measurement
The amount of uncertainty in reporting scores; the degree of imprecision that may result from the measurement process (e.g., test content, administration, scoring, or examinee conditions), thereby producing errors in the interpretation of student achievement. Technically, it is the typical variation between observed scores and theoretical true scores.

Equivalent Groups
Groups of test takers whose comparable characteristics (abilities, performance, reliability) appear to be essentially the same, having very similar (if not identical) effects.

Matrix Sampling
A measurement technique whereby a large set of test items is organized into a number of relatively short item sets or blocks. Each subset is then administered to a subsample of test takers, thus avoiding the need to administer all items to all examinees.

Order Effects
Fluctuations in scores that arise from the order in which test questions (or entire tests) are taken. For example, items in Test Section 1 may help prepare test takers for the items in Test Section 2, and vice-versa.

Parallel Tests
Two or more versions of a test considered to be interchangeable in that they are built to measure the same constructs, are intended for the same purposes, are administered using the same directions, and are designed to yield comparable scores. Also referred to as *alternate* test forms.

Random Sampling
The selection of a sample such that the selection of each element in is no way dependent on the selection of any other element.

Sample
A sample is a selection of a specified number of entities, called sampling units (test takers, items, etc.), from a larger specified set of all possible entities, called the population.

Spiraling
Refers to a way test booklets are packaged and distributed to students. In a spiraled test administration, test takers are randomly assigned different test forms, distributed in such a way that each test taker's form is different from the adjacent test taker's form.

End Notes – Chapter 3

[1] The definition of "large sample" can vary from as small as 200 to as large as 8,000 examinees, depending on the overall assessment program, the IRT model used, and other factors. Under the Rasch or 1PL model (see Chapter 2), 500 to 1,000 students might be considered sufficiently large; under the 3PL model, 3,000 to 5,000 students are often sufficient, although some researchers prefer numbers as high as 8,000 to 10,000. Readers interested in accessing research literature with regard to sample sizes are encouraged to review Downing and Haladyna (2006, pp. 495-496), Dorans, Pommerich, & Holland (2007), and Kolen & Brennan (2004, pp. 288-289). These texts discuss various sample sizes for a myriad of equating applications.

[2] Kolen & Brennan's 2004 text (*Test equating, scaling, and linking)* does not use the term "Anchor Test design" specifically but describes the same principle presented here—a common-item set, which is a "mini version" of the total test, is internally located within the test—under the Common-Item Nonequivalent Groups Design. However, in the more recent *Linking and Aligning Scores and Scales* (Dorans, et al., 2007), Kolen provides a description of Anchor-Test Nonequivalent Groups Design for Linking (p. 47). In short, Kolen explains that if two test forms are substantially different, the common items can't be a mini version that represents the whole of *both* test forms, and thus an anchor test is required. This is a relatively recent and subtle differentiation, as many practitioners may still informally accept the term "anchor test" to denote the common items used as anchors on two test forms.

[3] Some texts make the distinction between *internal* and *external* anchors without using the terms *embedded* and *appended.*

Chapter 4

The Mechanics of Equating

This section of the handbook illustrates the "how" aspects of equating by describing the equating tools used under CTT and IRT. The information is intended to acquaint readers with essential ideas that measurement professionals frequently encounter in large-scale assessment.

4-A. Conceptual Overview

In 1992, Mislevy described four typologies of linking test forms: *moderation, projection, calibration*, and *equating* (Mislevy, 1992, pp. 21-26). In his model, *moderation* is the weakest form of linking tests, while *equating* is considered the strongest type. Thus, equating is done to make scores as interchangeable as possible.

It is helpful to think of equating as part of a linking continuum, as shown in Figure 4.1. This figure shows equating as the strongest kind of linking, with all other weaker forms of linking found somewhere to the left.

The equating side of this continuum also represents more demanding assumptions that allow scores from two or more test forms to be used interchangeably. Links that cannot satisfy the strict requirements of equating may still be described as being *toward* the right of this continuum, but cannot claim to be equating.

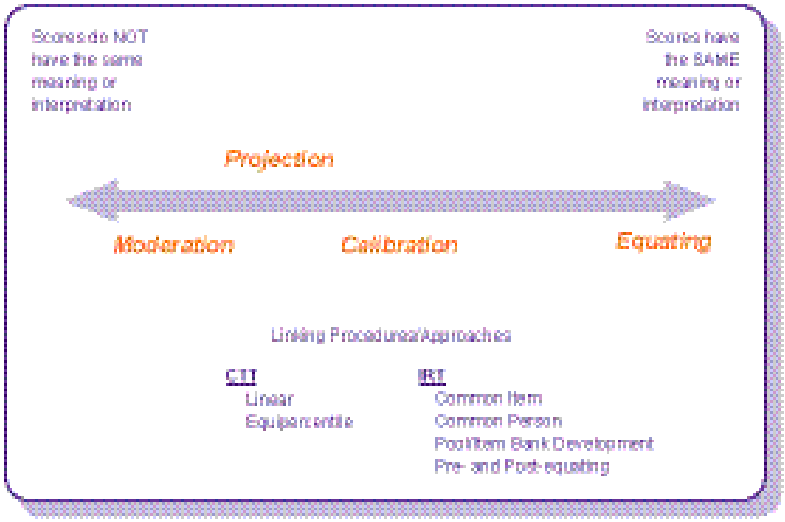


Figure 4.1: The Linking Continuum

Although practitioners sometimes refer to all forms of linking as equating practices, *equating* is more accurately described as *the most strict and demanding form of linking*. A thorough examination of the other typologies (moderation, projection, and calibration) will not be covered in this handbook; instead, this chapter will focus on the fundamental concepts and tools practitioners use for *equating*.

There are various tools or procedures for linking test forms, some associated with Classical Test Theory (CTT) and others with Item Response Theory (IRT). All of these procedures can be used for equating and other kinds of linking. However, it is essential to recognize that when used for *equating*, these procedures apply to tests that **have been constructed to be parallel** so that scores on multiple forms have the

same meaning or interpretation. There are strict technical definitions of "parallel" in this context but essentially it means test forms were constructed to measure the same content, at the same level of cognitive complexity, to have the same mix of item types (selected response, short answer, extended response) and to use the same test format (e.g., paper and pencil, computer based). When forms are designed to be parallel, the remaining task for equating is to create equivalence between scores, which allows scores for each form to be used interchangeably.

The following sections focus and expand upon the concepts of the linking continuum first described in Chapter 1. They provide the reader with a conceptual overview of the most commonly used tools and procedures employed to deal with the mechanics of equating.

4-B. Classical Test Theory (CTT) Equating

Linear Equating

Linear equating is a tool used primarily under CTT for determining equivalent scores between two parallel test forms. Linear equating is based on the assumption that two parallel test forms to be equated have similar **distributions** of scores except for their **mean scores** and their **standard deviations** (Crocker & Algina, 1986, p. 458).

Linear equating draws its name from the fact that the relationship between scores of Test A and Test B can be shown as a straight line on a graph. The line thus represents the equivalent relationship for all possible scores. In the example shown in Figure 4.2, Test B (a new 10-item test) appears to have more difficult items than Test A (an older 10-item test) based on the mean scores (Test A mean=7, and Test B mean=5). How would a score on Test B translate to a score on Test A? To find out, we can use linear equating to represent the relationship graphically:

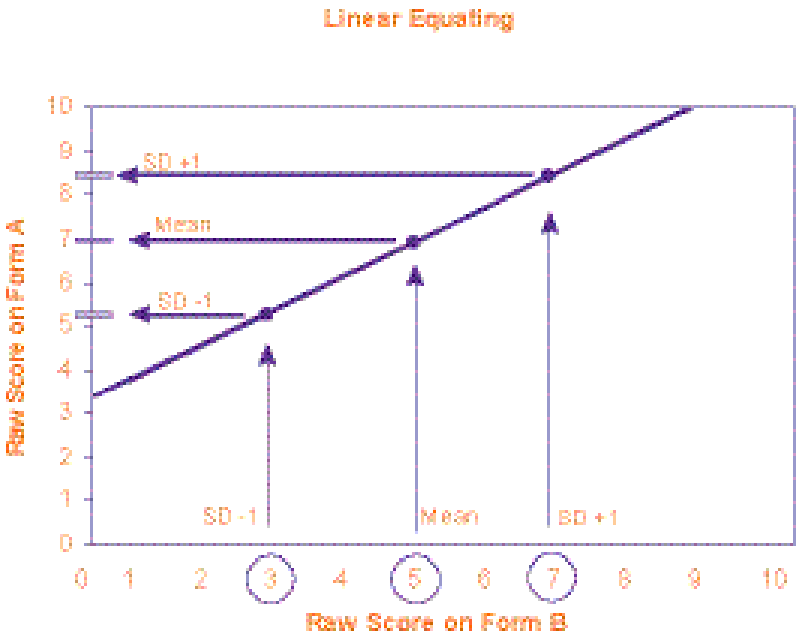


Figure 4.2: The basic concept of linear equating

Linear equating assumes that the only differences between the two tests are mean and variability (as measured by standard deviation or "SD"). In Figure 4.2, to equate Form B to Form A, we transform the mean score of Form B (5) to the mean

score of Form A (7). We then transform scores on Form B to scores on Form A for one standard deviation above and below the mean (3 and 7 in this example). The net result is a straight-line transformation of scores from Form B to the score scale of Form A.

The example shown in 4.2 is intended only to demonstrate the basic concept of linear equating. There are different approaches, depending on the assumptions made and the characteristics of the samples used.

Advantages/Disadvantages

Figure 4.2 is a simplified example of linear equating, but it also illustrates some of the limitations of this method. Notice that the scores for Form B (3 and 7) are discrete: they do not have decimal values. But they do not transform to discrete raw scores on Form A. For example, a score of 7 on Form B is the adjusted equivalent of the score 8.5 on Form A. However, students who take Form A cannot receive a raw score of 8.5! To deal with the problem of non-discrete equivalent scores, psychometricians use various approaches to round the equivalent scores in ways that allow them to report discrete scores. Rounding, however, introduces its own kind of equating error.

Notice also that Figure 4.2 shows a linear relationship in which a very high score on Form B results in a score on that is outside the range of possible scores for Form A. The graphic seems to suggest that a score of 10 on Form B will equate to a score of 11 or higher on Form A, which is not possible. (Most decisions about students with perfect or near perfect scores, however, are quite clear even if exact equated values cannot be found.) This is not an error in the way the graph is drawn; rather, it is part of the nature of linear equating. And although it is not necessarily a problem for students whose scores are near the mean, having a score on Form B that transforms to a score outside the range of possible scores for Form A can be difficult to explain (Livingston, 2004, p. 8). Additionally, when the two forms to be equated are very different in terms of difficulty level, the "shape" of the distribution of scores can be quite different for each form, causing the line of the graph to either be very steep or very flat. This characteristic of linear equating makes this method most appropriate when the accuracy of equating results is most important for scores that are near the mean (Kolen & Brennan, 2004, p. 293).

Linear equating is generally considered one of the easiest equating procedures to perform. It is relatively simple in terms of the mathematic calculations. This simplicity is often cited as an advantage of linear equating. However, modern advances in computer hardware and software have probably rendered the argument obsolete.

Like all equating methods discussed in this chapter, linear equating makes the assumption that the two test forms are parallel. In addition, as noted above, it requires key assumptions about the *similarity of the distributions of scores* for the two forms. In particular, it means that one key assumption holds true: the mean scores and score variances for the two distributions are the only significant difference between them. But what if one test form has a normal score distribution, while the other is very skewed? In these cases, the accuracy of this equating method may be called into question—especially for students whose abilities could be described as very strong or very weak. Other equating methods, such as equipercentile equating, are designed to handle greater differences in test difficulty than linear equating. They provide an alternate method of transforming scores from Form B to Form A, especially for cases where examinee groups are either very strong or very weak in terms of the construct being assessed.

Equipercentile Equating

The equipercentile method provides for accuracy of equating results along the entire score scale. It also allows for more accuracy than linear equating when test forms differ in overall difficulty level (Kolen & Brennan, p. 294).

The first step in equipercentile equating is to determine the **percentile ranks** for the score distributions of each of the two tests to be equated (Crocker & Algina, p. 462). Percentile ranks between the two test forms are then "equalized."

To illustrate, suppose that two 10-item forms (A and B) of a mathematics test are to be equated using this method. As was the case with linear equating, it can be assumed that Form B has been constructed to be a parallel version of Form A.

After students take the two forms, percentile ranks are computed from raw scores. Then, raw scores can be paired with percentile ranks, as shown in Figure 4.3:

	Percentile Ranks (by Test)	
Raw Score on Hypothetical 10-Item Math Test	Test Form A	Test Form B
10	99	99
9	95	93
8	90	**80**
7	**80**	62
6	65	42
5	45	**25**
4	**25**	15
3	13	10
2	5	5
1	1	1

Figure 4.3: The basic concept of pairing percentile ranks

The table in Figure 4.3 shows that if students get 7 out of 10 questions right on Form A, their score is at the 80th percentile for the group taking the test. However, when students take Form B they must answer 8 of 10 questions correctly to reach the 80th percentile. Why? Because Form B appears to be harder than Form A.

Note that in the table above, students in the 25th percentile on Form A make a score of 4, whereas students in 25th percentile on Form B make a score of 5. A score of 4 on Form A is therefore equated to a score of 5 on Form B, because these scores represent equal percentiles on each test (hence the term *equipercentile*).

The equipercentile equating method uses tables like the one shown above as a basis for deriving equivalent scores from percentile rankings for Form A to Form B. In practice, tests contain many more test items than the simplified example above.

In addition to using tables, we can also generate equivalent scores for percentile rankings by plotting scores onto a graph, as illustrated in Figure 4.4:

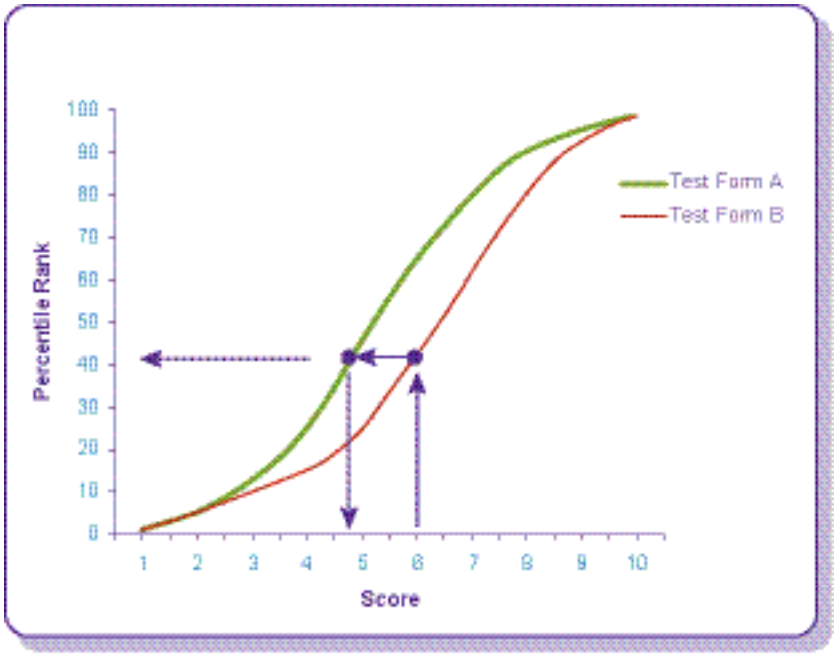


Figure 4.4: The basic concept of equipercentile equating

In the example above, percentile rankings are plotted for scores of two hypothetical 10-item tests (Forms A and B). Comparing the percentile rankings graphically is another way to demonstrate the concept of equipercentile equating. The score of 6 on Form B results in a percentile ranking of just over 40; in contrast, a score of 4.75 on Form A is required to achieve the same percentile rank.

Percentile rankings can be important for reporting purposes if percentiles are used for communicating the results of tests to students and parents (Crocker & Algina, p. 440). Computer software can use the concepts illustrated above to calculate equivalent scores and percentile rankings for all possible raw scores obtained on Forms A and B. (In practice, raw scores might be rounded up or down since no student could actually receive 4.75 as a raw score on Form A.) So for the simplified example above, the equipercentile method of equating could support an interpretation such as "a person who earns a score of 6 on Form B would most probably have earned a score of 5 on Form A, and the student's score ranking for either test can be reported at the 42nd percentile."

When using the equipercentile method, the distributions of the scores on both tests are often somewhat irregular; instead of the smooth distribution of scores represented by the hypothetical bell curve, scores distribution curves look "bumpy." In these cases, smoothing techniques are useful for addressing this problem. Several different smoothing methods are available, and you should discuss these methods with your contractor if equipercentile equating is being considered.

Advantages/Disadvantages

The equipercentile method equates scores that are within the range of possible scores for both tests, which overcomes one of the problems with linear equating (Kolen & Brennan, p. 47). The equipercentile method also carries fewer assumptions than the linear method about differences between the distributions of scores for Form A as compared to Form B. On the other hand, **equating error** is generally larger with this method than with linear equating (Crocker & Algina, p. 465).

Linear vs. Equipercentile Equating

Equivalent scores are the outcome of both the linear and equipercentile equating methods. Scores for Form A can be considered as having equivalent scores on Form B if both tests "measure the same trait with equal reliability and the percentile ranks corresponding to the scores are equal" (Crocker & Algina, p. 457). Because of many similarities between these two methods, some experts consider linear equating to be an approximation of equipercentile equating (Hambleton et al., 1991, pp. 124-125).

Many factors may justify the use of one method over another, but one primary consideration is how well the assumptions of linear equating hold up—namely, that the tests to be equated differ only in terms of their mean and variability (as measured by standard deviations). In contrast to linear equating, the equipercentile method makes fewer assumptions, so the equipercentile method is likely to be more accurate when assumptions of linear equating are not tenable. The equipercentile method is considered *less accurate* than the linear method in cases where score distributions are not too different (Crocker & Algina, p. 465), although equipercentile and linear methods will produce similar results if the distributions differ only in the mean and variability.

4-C. Item Response Theory (IRT) Equating

Overview

The use of IRT models and methods allow for considerable flexibility in equating tests and building equivalent test scores. This section provides an overview of the key concepts and methods at a conceptual level. The examples intentionally use cases with a small number of items in order to illustrate key ideas and principles. A number of examples are illustrated with the 1-parameter IRT model to minimize the discussion of psychometric details. In practice, many large-scale assessment programs use this model, but many programs are also supported effectively with the 2- and 3-parameter IRT models.

Two general approaches to equating using IRT are observed score equating and true score equating (both are also used under CTT). Observed score equating directly connects the scores on one test to another, while true score equating involves estimating and connecting the true scores (described in Chapter 2) on two test forms to determine the relationship between observed scores.

Kolen & Brennan (2004, p. 185) describe the outcomes between observed score equating and true score equating as producing similar results in some cases but somewhat different results in others. It is important to understand which procedure your test developer uses and the rationale for choosing the procedure.

Equating through Common Items

Equating with IRT methods falls into two broad categories: equating through common *items* and equating through common *people*. The following sections cover the most typical applications of common *item* equating:

- equating by applying an equating constant
- equating by concurrent or simultaneous calibration
- equating with common items through test characteristics curves

Equating by Applying an Equating Constant

Common item equating is done by using common items embedded in two different test forms. To illustrate, the following explanation will use the simplest IRT equating method and the simplest IRT model (the 1-parameter, or *Rasch*, model), to demonstrate a set of important basic principles of IRT equating. By extension and variation, these principles are applicable in most IRT equating situations. Other IRT models and other situations may allow for greater precision in equating than would be obtained in this example.

The first step in any equating analysis involves assigning difficulty values for all items, including the common items. This is done using a statistical procedure called *difficulty estimation*. There are many approaches and options available when performing a difficulty estimation, but the first example presented below will show an estimation done in such a way that the average of *all* difficulties for *all* the items is set at 0. Although any other value for fixing the origin for the scale can be chosen, setting the average of the item difficulties to 0 is useful because all easier-than-average items have *negative* difficulties and all harder-than-average items have *positive* difficulties.

In most IRT applications, the origin or starting point of the measurement scale is critical and can generally be set at almost any convenient point. Sometimes the first test form of an annual program is used to define the origin; in others the ability needed to reach some performance level (e.g., "proficient") might be used as the origin, or the mean of the students' abilities might be used. Scales can be shifted quite easily to use different origins, which we will see. But first we begin with an example shown previously in Chapter 2: the easy-to-hard item difficulty continuum.

In Figure 4.5 below, for example, Form X contains 20 total items—three of these are identified as **anchor items**, which are defined as common items used to link test forms—and all 20 items are placed on the continuum. Note that this simple example uses three anchor items, but in practice at least 15–20 items would typically be needed as anchors.

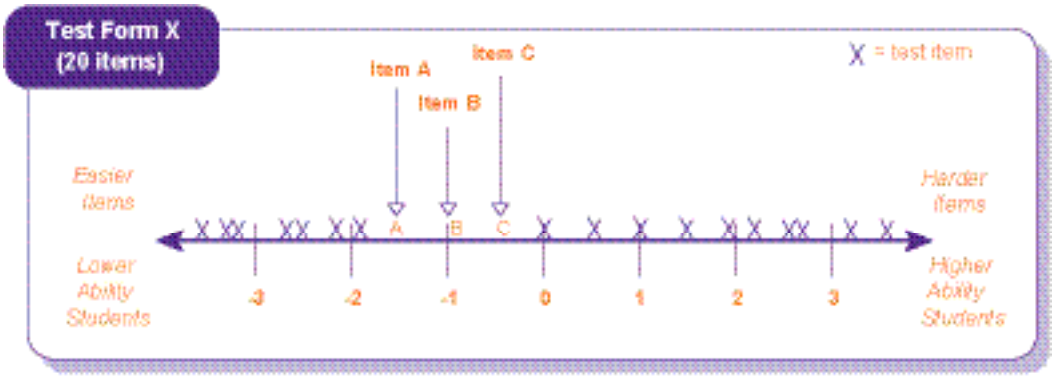


Figure 4.5: Illustration of three relatively easy anchor items
In this illustration, items A, B, and C are identified as anchor items. Each anchor item has a less-than-average level of difficulty (average difficulty for all 20 items is zero.)

In Figure 4.5, the anchor items are relatively easy compared to the origin of 0. More specifically, they are collectively easier than the average item difficulty by -1. This means that, on average, the other non-anchor items on Form X are more difficult than the A, B, C anchor set.

Now consider Figure 4.6, which shows the use of anchor items A, B, and C, but in this example the 17 non-anchor items on Form Y are different:

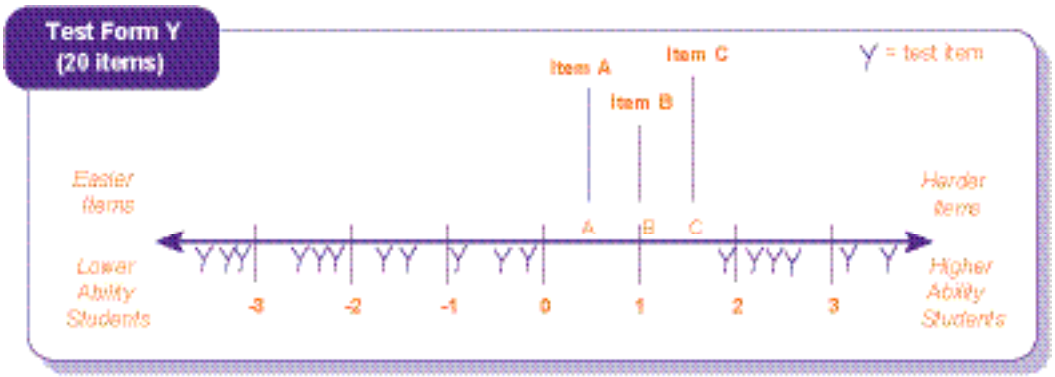


Figure 4.6: Illustration of three relatively difficult anchor items
In this illustration, items A, B, and C are also identified as anchor items, but each anchor item has a greater-than-average level of difficulty (average difficulty for all 20 items is 0.)

In Figure 4.6, the same anchor items appear to be relatively difficult. More specifically, as a group they are harder than the average item difficulty by +1. This simply means that the other non-anchor items on Form Y are, on average, easier than the "A, B, C" anchor set.

The key to equating items on Forms X and Y is to observe that the average difficulty of the anchor items shifts from -1 for Form X to +1 for Form Y relative to the origin of 0. This is a shift or difference of 2 (the difference between -1 and +1). This shift or adjustment is known as "applying an equating constant." It can be represented visually by "lining up" the anchor items of Figures 4.5 and 4.6:

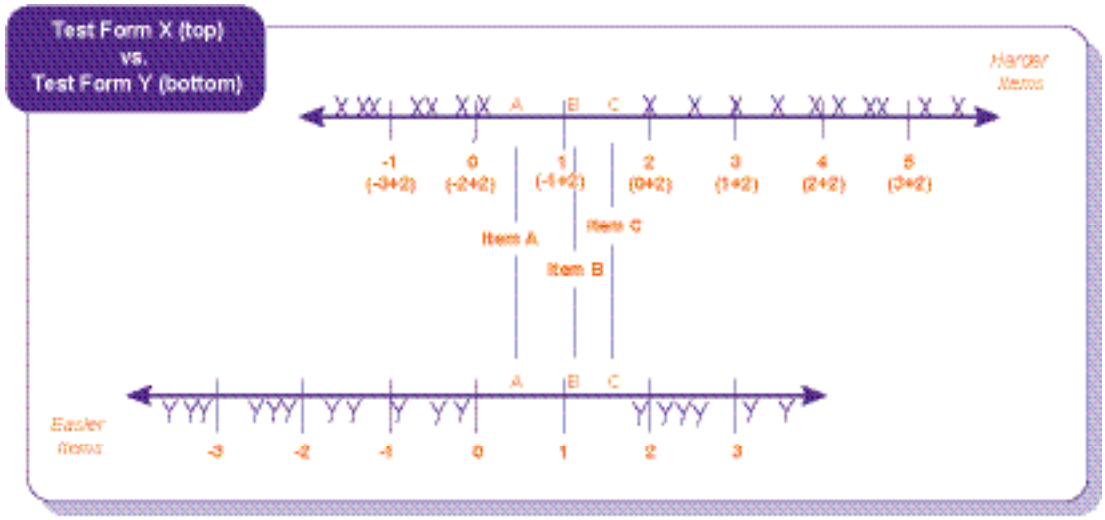


Figure 4.7: "Lining up" anchor items
Although anchors A, B, and C were relatively easy in Test Form X, they were harder items in Test Form Y. "Lining up" the anchor items (and finding their differences in difficulty from one test to another) can be the first step toward placing items on the same scale, thus providing "equal" values between the two tests.

If we take the value of this shift (2) and add it to the difficulty of all items on Form X, the result is that the average of the anchor items on the adjusted Form X becomes +1, the same as their average value on Form Y. Another way of thinking about the shift is to note that in Figure 4.7, the anchor items are "lined up" by adding the value 2 to each point of the scale for Form X so that the result is a scale for Form X that is the same as Form Y: -3 becomes -1, 0 becomes 2, et cetera.

We can also see that by making this adjustment to Form X, the average difficulty of the anchor items on the adjusted Form X* would be +1, which are equal to (i.e., equated to) the difficulty of the anchor items on Form Y. More importantly, all of the items on Form X can be placed on the same scale as all of the items on Form Y, which is shown in Figure 4.8:

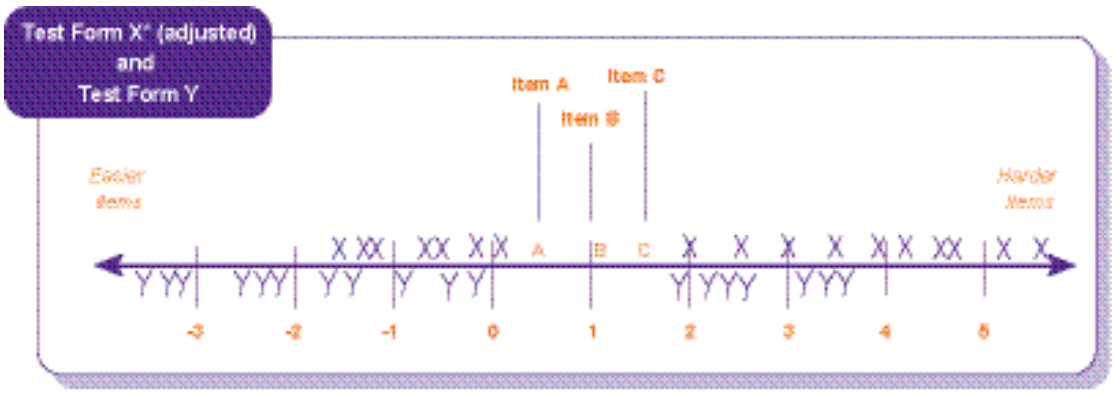


Figure 4.8: Two test forms placed on the same scale via anchor items
By determining the difference between the average difficulty of anchor items on Form X versus Form Y, Form X can be adjusted to Form Y.

In this illustration, Form Y has been fixed and defines the origin on the scale. Form X has been shifted by +2 units so that the averages of the anchor items' difficulties are lined up, i.e., they are equal or equated. The result is that Forms X and Y are now equated, and all of the items on the two test forms are now on the same scale. This

includes 17 items unique to Form Y, 3 anchor items common to both forms, and 17 items unique to Form X—a total of 37 items.

The process shown for linking Test Forms X and Y could be repeated to link many tests, constructing a bank of items—all of which would be on a common scale. Likewise, the proportion of anchor items (3) to form-specific items (17), as shown in Figure 4.5, could be reversed to have multiple versions or forms of a test (perhaps 10 versions) with a large number of common items (perhaps 50 items) and a relatively small number of unique items per test version (e.g., 10 items). This design could be used to put the 10 sets of 10 items—a total of 100 items—onto the same scale as the original 50 items.

Reflection on the Example

Figures 4.5–4.8 provide a highly simplified illustration intended to show the basic ideas and procedures used in common item equating. Other important qualifications to note in connecting this example to more practical applications include the following:

- Test forms cannot be equated with only three items. A magic number or proportion of common items is hard to specify, but a minimum of 15–20 items is common practice for test forms containing approximately 40–60 items.
- Test forms that vary in difficulty as much as Test Forms X and Y in the example would be equated or linked only under special circumstances. Such circumstances might include cross-grade (vertical) linking/equating, or equating for a very wide range of abilities.
- Test forms constructed to use anchors/common items should use common items that span the difficulty of the overall test much more broadly than in the example: anchor sets should contain easy, moderate, *and* hard items whenever possible. In addition, anchor items should reflect overall content and item formats used on the full test form.
- The basic common item approach can be used with short answer or extended response items scored with ordered values from 0, 1, 2, 3, 4, etc.

Some tests items may work well on Test Form X and well on Test Form Y, but do not function effectively as anchor items in trying to equate test forms. (For example, using only very easy items or very hard items as anchors may yield problematic results.) There are a number of procedures available to test the adequacy of items as linking items. Practitioners should insist that contractors evaluate items for their suitability and stability as linking items.

Equating by Concurrent or Simultaneous Calibration

Another common procedure used with all IRT models involves identifying the common items on a test form or item bank and using them as the origin for equating *without* adding (or subtracting) an equating constant. Instead, when using the concurrent or simultaneous calibration approach, the IRT item characteristics for the common items on one test form are considered fixed or anchored. In a certain sense, the IRT difficulty, discrimination, and guessing values for the items are treated as if they are true values and they are not allowed to vary.

For example, consider a case in which the IRT item difficulties from Form X are considered fixed and anchored. Form Y can be equated to Form X when the items on Form Y are scaled or calibrated to obtain their IRT difficulties. In this special scaling or calibration of Form Y, the IRT values for anchor items (A, B, and C) are fixed to their Form X values and all the items unique to Form Y are scaled or calibrated so that they are forced onto the scale defined by the values of these fixed anchor items.

This concurrent or simultaneous calibration approach has considerable flexibility and can be used to equate multiple forms simultaneously.

Equating With Common Items through Test Characteristics Curves

A third procedure for equating test forms that share common items is based on the work of Stocking & Lord (1983) and is used quite frequently with data analyzed using the 2- and 3-parameter IRT models. The basic logic of this approach works quite well in a wide range of settings. Like all equating procedures, it assumes that the test forms involved were built to be parallel.

The Stocking & Lord approach works through IRT Test Characteristics Curves (TCCs). A TCC shows the relationship between the IRT ability and the expected raw score on a test. Students with higher abilities are expected to have higher raw scores than students with lower abilities, and the relationship shows the logistic shape of test forms, which have a shape similar to item characteristic curves:

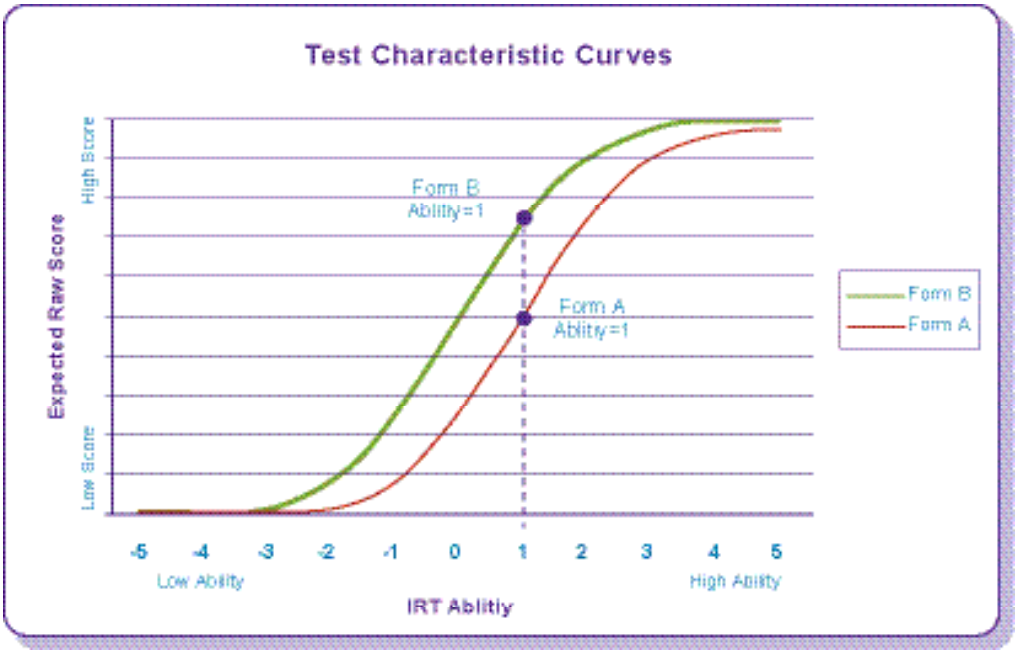


Figure 4.9: Test Characteristics Curves

In the example shown in Figure 4.9, these two test forms share common items, but each form also has its own set of unique items. Test Form A, the test to the right, is shown as more difficult since the same IRT ability position (on the horizontal axis) maps into a lower raw score than the raw score on Test Form B that corresponds to the same IRT ability.

The Stocking & Lord approach is quite flexible and is used in a wide variety of equating situations. To use this procedure, a choice of origin must be made. Practitioners must decide if Form A will be the origin and stay fixed, with Form B equated to the Form A scale; alternately, Form B could be the origin and stay fixed,

with Form A equated to the Form B scale. For this example, we'll assume that the Form B scale is being equated to the Form A origin. (For the sake of completeness we should mention that any average of Forms A and B could be used as the origin or they could both be equated to some third form or a preexisting item bank scale).

The procedure then uses the parameter estimates of the common items along with a series of steps used to find weightings that can be applied to the item parameters for the common items on Form B. These weightings make the discrimination, difficulty, and pseudo-guessing parameters of the common items on Form B as close as possible to the corresponding values of the same common items on Form A. When weights are applied to all Form B item parameters, Form B is equated to the Form A scale. The success of the procedure is evaluated by comparing the TCC for the common items for the examinees who took Form B to the TCC for the common items for the examinees who took Form A at several points along the IRT ability range. This is a very condensed and simplified explanation of the Stocking & Lord procedure, and there are many variations and technical details described in the original article and subsequent publications. For example, test forms could be equated to a hypothetical TCC constructed from items from an item bank.

Common Person IRT Calibration

The basic logic of linking or equating by fixing and holding constant the difficulty of a set of common items can also be applied, in reciprocal fashion, to common people. Thus, common *person* IRT calibration might be considered the obverse of common item equating: instead of using previously tested *items* with previously fixed parameters, we use previously tested *people* with previously fixed abilities. In other words, to perform common person equating, the IRT ability of students is estimated based on one test. These ability estimates are then fixed and held constant when students take a different test.

By analogy, first suppose that 5,000 students took a 50-item test and were measured on an IRT scale. (Recall that in Chapter 2, the basic concept of scaling under IRT showed both students and items on the same scale.) If these same 5,000 students were then given a new set of 30 items, we could use the IRT scale values for these *people* to place the new set of 30 items on the same scale used to measure these students. The 50-item test and the 30-item test have no items in common, but these two tests do share something that can be used to compare them: they share the same 5,000 test takers, whose abilities are known.

Consider a different analogy: suppose you have a sample of 100 people and you know with certainty each person's height in inches. Suppose also that you have a tape measure with uniform markings, but no actual numbers. In this hypothetical situation, you could simply use the 100 people who were previously measured to assign scaled values to the "numberless" tape measure. The height of each person is fixed—we merely use the measurements we know to assign scale values (i.e., mark the tape measure) accordingly.

Pre-Equating and Post-Equating

Pre-equating

The use of **banked items** that have established IRT item parameter estimates along with embedded field test items, which have no previously determined or fixed parameter estimates, allow psychometricians to estimate IRT parameter values for the field test items after administration. This is possible because the field test items become attached to the same scale as the fixed-value bank items. Field test items that perform well (meaning their statistical features are appropriate) are added to the item bank and thus become candidates for pre-equated test forms.

Pre-equated test forms are operational test forms constructed from item banks. Each test question carries its own IRT item parameter estimates from the bank, thus allowing developers valuable flexibility in terms of selecting a set of items that are likely to prove viable for operational testing.

The primary goal of item pre-equating (and its chief benefit) is the ability to produce raw-to-scale score conversion tables *before* a form is administered (Kolen & Brennan, 2004). As test users and developers face increased pressure to deliver test results quickly, the practice of preequating test forms has also increased. Although the results of preequating procedures have generally been considered acceptable (Taherbhai & Young, 2004; Bejar & Wingersky, 1982), these results suggest parameters for some items do change or drift from one test administration to the next (see *item parameter drift*). Therefore, states using pre-equating techniques should build in procedures to update item parameters over time.

Position effects are an important consideration when using banked items for pre-equated test forms. For example, items that are initially field tested as questions 50 through 55 on a field test form may show significant differences in terms of the **item parameter values** when these same items are moved to earlier positions on an operational test. It is important to discuss the role of item positioning effects with your contractor, and the controls in place to manage these effects, when pre-equating is employed. (Also see the anchor items section of Chapter 5 for further discussion of anchor item positioning).

Post-equating

Post-equating is commonly used to offset the influences of the field-testing situation on item parameters. For example, if students are aware that a field test does not "count" and see the administration of the test as inconsequential, item parameters may be unstable and change once the field test items are used on operational test forms. Pre-equated forms constructed based on field test data may prove to be more or less difficult or show other variations when compared to how these forms perform in an actual operational administration.

When time and resources allow, it is desirable to use pre-equated item information to construct operational forms and then rescale the forms using data from the operational administration. **Rescaling** is often done using a carefully selected "early return sample" that includes schools with students who, when taken together, are representative of the state population. This process begins when the early return sample schools return answer documents on an accelerated schedule for early scoring. Then the data are analyzed, scaled, and equated using the procedures described in this chapter. Thus, the early return sample allows psychometricians to ensure that the "live" test score distributions are employed, and, using this information they can make certain adjustments to account for variations as part of the post-equating process.

A final step in this sequence of pre- and post-equating often involves updating the bank values. In this final step, the post-equated IRT parameters of the items used on the operational test form (or forms) are fixed and used as the new values for the items in the item bank. The item parameters of the field tested items that were not used in the first operational forms are then equated to the new origin defined by the operational test post-equating. Furthermore, in practice, each time an item is administrated, its IRT parameters should be updated as part of a standard procedure for maintaining item banks (see Chapter 5 section on item banking).

Chapter Glossary:

Anchor Items
A set of items common on two or more test forms that are used for the purpose of linking/equating test forms.

Banked Items
Items previously field tested that are part of a secure item bank and available for future test construction if item content and IRT parameters are suitable for this purpose. (See also item bank development, Chapters 3 and 5.)

Distribution of Scores
The number of examinees at each score level.

Equating Error
The random error inherent in the equating process or procedure.

Item Parameter Drift
An effect that occurs when newly constructed test forms are administered and IRT parameter values (such as item difficulty, discrimination, and pseudo-guessing) are found to differ significantly from the previously established values of the item bank.

Mean Score
The score computed by totaling the scores of all examinees, then dividing that sum by the number of examinees.

Parallel Test Forms
Two or more test forms built to be equivalent in terms of content, cognitive demands, and item format.

Percentile Rank
Technically defined as the percentage of examinees with lower scores, plus half the percent with the same exact score. Also defined more generally as "the percentage of examinees with the same score or lower scores."

Rescaling
The process of **scaling** previously scaled raw scores available from a more recent test administration; often performed on an early-return sample of test results as part of a post-equating. (Post-equating is often a necessary or required quality control check intended to verify the viability of previously pre-equated forms.)

Scaling
The process of associating numbers or other ordered indicators with the performance of individual test takers. Raw scores are transformed to *scale scores* using statistical methods. Typically, *scales* are constructed in ways that will help test users interpret the scores.

Standard Deviation
A measure of the variation present in a distribution of scores; sometimes interpreted as the average amount that scores in a distribution of scores deviate (differ) on both sides of the mean.

Chapter 5

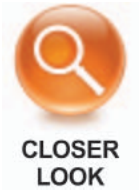
Common Equating Issues

This section of the handbook deals with common equating-related issues encountered in statewide testing programs and some other large-scale assessments. It draws upon the experience of measurement professionals who have faced these equating situations with different approaches and solutions. Certain topics discussed here have been addressed earlier in this document and are repeated here for the sake of completeness.

Each subsection of this chapter contains four components:

1. a question that represents a common issue or concern
2. a brief explanation of the issue
3. some key ideas to consider and/or possible ways of handling the issue
4. key questions to ask assessment contractors about the issues

5-A. Changes in Test Specifications



Changes in Test Specifications

COMMON QUESTION OR ISSUE

Q: What happens if test specifications change?

EXPLANATION

Assessment programs evolve as educational assessments attempt to become more responsive to legislative, political, and practical demands. Changes in test specifications due to shifts in curricular content and adjustments in educational policy present challenges to equating and the ensuing interpretation of results.

KEY IDEAS TO CONSIDER

Test specifications need to remain relatively stable from one year to the next in order for **equating** to work well. Per Kolen and Brennan, "equating can be successful only if the test specifications are well-defined and stable" (p. 270). If test forms are significantly different from one year to the next, the point at which test scores can be considered interchangeable is called into question.

Minor changes in testing programs can generally be accommodated with minimal difficulty (Kolen, p. 272), but major changes often make equating very challenging and may require a radical overhaul and updating of the item bank and the scaling process.

In practice, the stable portion of a test, i.e., the subset of items that represents content that has been used and assessed before, can be treated as an anchor set. The degree to which the items measuring new or modified content standards or changes in specifications can be constructed or fit on the same scale should be evaluated. The decision about whether changes in specifications *support* or *work against* equating should not be based on assumptions, but informed by empirical tests of dimensionality and data-model fit.

KEY QUESTIONS TO ASK

How much latitude might curriculum consultants have to suggest alterations to test specifications without jeopardizing the technical defensibility of the equating?

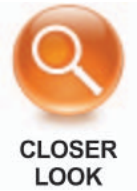
What efforts are made to ensure similar test specifications are used on new forms in terms of the number, type, and content of the items?

What tests of data-model fit and trait stability over forms have been performed?

When does a change in test specifications become significant in terms of the time and effort required to adjust for the change?

When changes are deemed significant, what is the result? How is the assessment program overall impacted by the change?

5-B. Anchor Item Considerations



Anchor Item Issues

COMMON QUESTION OR ISSUE

Q: What are the most common issues involved in using anchor items as part of an equating design?

EXPLANATION

Anchor items play an important role in many equating designs because they are the basis for equating and accurate equating helps determine performance changes across years or across forms. Issues concerning anchor items are common. Several factors may affect an item's suitability for use as an anchor item:

- content representation
- item format representation
- differences in adjacent or nearby items that might clue the answer to an anchor item
- item parameter drift
- compromised security of a test item resulting in performance differences for an anchor item
- content "discoveries" (such as the demotion of Pluto to non-planet status) that result in answer key changes
- test takers seeing an anchor item on a previous test (or tests)

KEY IDEAS TO CONSIDER

Most test publishers have internal controls in place to help identify clueing or context effects that may be present when using pretested anchor items with untested items.

To check the behavior of anchor items, IRT or classical item statistics can be used to examine items that behave differently from year to year or form to form within a year. There are a number of decisions to be made if anchor items are affected or if anchor item parameters are found to be unstable (parameter drift). First, if a small number or small proportion of anchor items seems unstable, there may be no impact in using the full set of anchor items. Second, anchor items with parameter drift that exceeds some fixed value or statistical criterion might be removed from the set of *anchor* items but still retained on the test and treated as a regular (non-anchor) item or a field test item. And third, anchor items might be weighted in the calculation of equating constants to reflect their respective instability (Cohn, Jiang, & Yu, 2008). In any case, removing anchor items may result in inadequate or imbalanced content distribution of the remaining anchor items, an inadequate number of items available for use as anchor items, or an impact upon previously established raw score cut scores.

Contractors should be required to document the procedures used to evaluate the stability of anchor item parameters and the results of applying these procedures.

RULE OF THUMB

How many items should be designated for an anchor test? Kolen and Brennan note that "experience suggests the rule of thumb that a common [Anchor] item set should be at least 20 [percent] of the length of a total test containing 40 or more items, unless the test is very long, in which case 30 common items might suffice" (Kolen, p. 271). Hambleton et al. (1991, p. 135) suggests a similar ratio (20–25 percent). In practice, using 15 to 20 items is not uncommon, but the appropriate minimal number depends on which IRT model(s) and which equating method(s) are employed.

An example of a common-item set that satisfies this rule of thumb is shown in Figure 5.1:

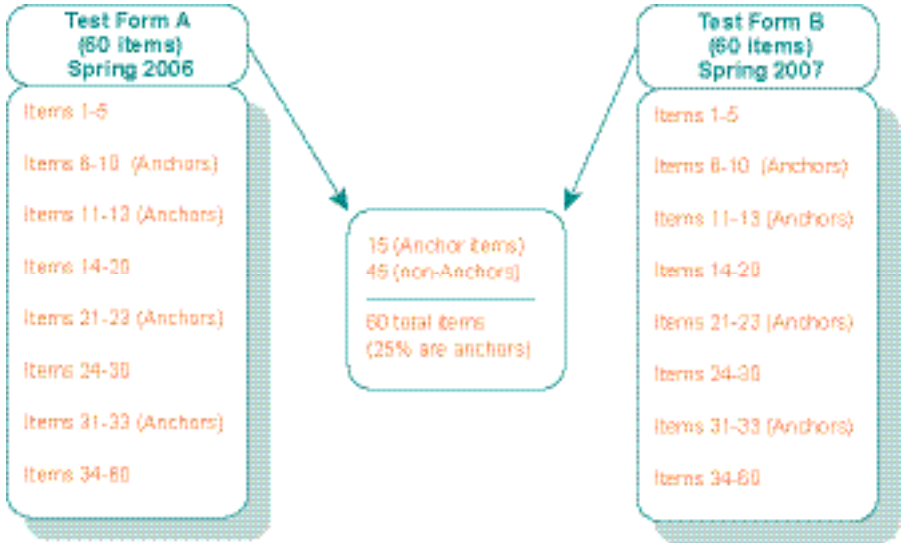


Figure 5.1: Illustration of the anchor item "rule of thumb"

The relative position of anchor items between two tests is also important. If an anchor item in Test Form A appears near the beginning of the test but is positioned as one of the last items of Test Form B, the difference in item position may be great enough to affect student performance on the item. Therefore, anchor items are often specified to occur within a fixed number of positions between the two tests. For example, if the tolerance for item positioning is set to within three items, an anchor item appearing as Item 10 in Test Form A could not be placed any higher than Item 13 in Test Form B.

RULE OF THUMB

The general rule of thumb for anchor item positioning is "the closer, the better." Anchor position difference tolerances are usually a function of the overall length of the test (and the psychometricians who specify the anchoring requirements). Generally, the anchor test equating method suffers when positioning rules are not followed closely.

KEY QUESTIONS TO ASK

What are the anchor test specifications/requirements overall?

What are the practical implications of using an anchor test design? For example, how might a test security breach or a misprinted item affect the equating results? What is the protocol for dealing with these kinds of issues?

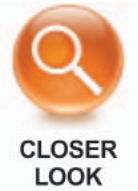
What suggestions are there for using groups of items associated with the same stimulus material as anchor items, e.g., items related to the same reading passage?

What is the tolerance for placement of anchor items from one form to the next? For example, does each anchor item need to be placed in the exact position on both tests, or will occupying a position in the same third of the test form suffice? Why?

How might changes in content specifications or test administration affect the suitability of the anchor set?

What controls are in place to ensure that test developers will utilize an anchor test that is suitable in terms of item quality and quantity? Are these controls documented?

5-C. Open-Ended or Constructed Response Items



Open-Ended or Constructed Response Items

COMMON QUESTION OR ISSUE

Q: Can equating be done on test forms that include the use of constructed response formats?

EXPLANATION

Open-ended questions refer to item formats that call for a short answer or more extended response. These item types are often referred to as constructed response formats. (See the following section on the special case of direct writing assessments.) These items are typically scored by judges according to pre-developed scoring **rubrics**, although the use of automated software to score responses continues to emerge.

KEY IDEAS TO CONSIDER

Notwithstanding technical cautions, many states routinely equate test forms that have a mixture of item formats. In most instances, the majority of the items used in equating are the multiple choice format. This equating strategy is common for equating tests within a grade across years when the mixture of formats remains substantially the same over time.

Kolen & Brennan (2004, p. 320) note three characteristics of open-ended items can make equating with constructed response items complicated: 1.) the rating or judging procedure, which is a source of error; 2.) the small number of items due to the longer per-item administration times; and 3.) the inapplicability of some equating designs or procedures because of too few items, administrative limitations that prevent spiraling of forms, or differences in scoring judgments that cannot be controlled.

KEY QUESTIONS TO ASK

Are judges or raters trained to reduce error? What methods are used to control for variation among judges? Are there any statistical procedures used to adjust for rater effects?

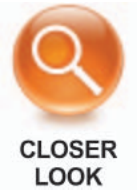
Can judges rerate papers from a previous administration to check on systematic rater drift over time?

What is the strategy for incorporating these items into the equating process? Will they be included at all?

Do items with the different formats fit the IRT model being employed? Are the various item types represented in the subset of items used to do the equating?

If using IRT, are there enough items present to allow for stable IRT calibration? Is the case count or sample size large enough for stable estimates? Does the IRT model to be used allow for a mix of item formats, including items that allow for partial credit?

5-D. Writing Assessments



Writing Assessments

COMMON QUESTION OR ISSUE

Q: Is it possible to equate test forms for direct writing? Why or why not?

EXPLANATION

Direct writing presents challenges to equating. In nearly every case direct writing assessments contain only one or two items (prompts). Writing samples might be scored for multiple traits by more than one rater but such assessments still provide limited data.

In addition, writing performance tasks are typically scored using rating scales that give only a narrow scale or limited range of raw score points. Consider a writing assessment where students are asked to respond to two prompts and each response (essay) is evaluated using a 0–6 scoring **rubric**. Such an assessment would be analogous to having a reading assessment based on only two passages, with each passage being associated with six selected response items. Both assessments (the writing and reading) would have a possible score range of only 0 to 12.

As a result, the small number of items and score points available on direct writing assessments make them difficult to analyze using IRT models, and can produce equating results which are difficult to explain.

KEY IDEAS TO CONSIDER

Some assessment programs have adopted writing assessments that include indirect writing measurements such as selected response items. This approach can considerably increase the total number of score points and thereby make some carefully monitored equating approaches possible. If these indirect measures focus on writing mechanics, their inclusion on the assessment may affect the construct validity of the writing assessment depending on whether the construct was intended to include writing mechanics. The addition of indirect measures focused on mechanics might undermine the validity of the writing assessment if the construct was intended to focus on composition and presupposes mechanics as an enabling objective.

In some cases, preliminary equating results might be such that there is decision not to equate because scores will be similar enough without equating. Rigorous training in monitoring of scorers may be adequate to support the claim of score comparability across prompts.

Measurement professionals may recommend that the equating of direct writing assessments be accomplished with the general polytomous form of the Rasch model, but various options/issues may be discussed (e.g., linear equating, mean equating, etc.).

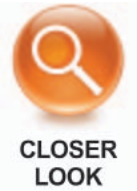
KEY QUESTIONS TO ASK

How will the inclusion/exclusion of writing tasks from the equating impact the overall assessment program?

What are the accountability implications of the decision to equate/not equate? What are the instructional impacts?

How rigorous are quality control procedures in the scoring process?

5-E. Paper-and-Pencil and Computerized Testing



Paper-and-Pencil and Computerized Assessments

COMMON QUESTION OR ISSUE

Q: Is it possible to equate computer-based assessments in the same way pencil-and-paper test forms are equated? How comparable are these two modes of administration?

EXPLANATION

The move toward computer-based assessment carries clear advantages, particularly for assessment contractors. Several studies have been done to compare paper-and-pencil tests to computer-based tests (CBTs) that present the exact same items in the same order as the paper-and-pencil test. Linking test scores between these two modes is usually undertaken to ultimately link test forms and may satisfy many of the strict requirements of equating (Dorans, Pommerich, & Holland, 2007, p. 137).

Another form of computerized assessment known as computer-adaptive testing (CAT) may not always support the strict requirements of equating to produce interchangeable scores, according to Dorans, et al. (pp. 137-139). In many cases, however, CAT scores and scores from paper-and-pencil tests can be solidly linked, if not technically equated in the strict sense.

KEY IDEAS TO CONSIDER

Developing a computer-based version of a paper-and-pencil assessment usually requires a special study to link scores the two modes of administration.

Random Groups, Single-Group with Counterbalance, and Anchor Test designs have all been used to create linkages, but more issues surface when using these designs to link CATs to paper-and-pencil tests, as compared to linear CBTs (Dorans, et al, p. 158). The linking between scores on a CAT and paper-and-pencil test will often present challenges in interpretations.

The comparability of various test formats may change over content areas and grades. Furthermore, the standards (AERA, APA, & NCME, 1999) require demonstration of equivalence across subgroups, not just all students in a particular grade.

KEY QUESTIONS TO ASK

What special comparability or linking studies would need to be done to integrate CBTs into the assessment program?

What are the requirements for implementing CBTs in terms of ongoing commitments of time, effort, and resources?

What kinds of linking among CBTs (and CATs) and paper-and-pencil assessments does the *current* body of research support?

5-F. Issues in Vertical Scaling vs. Horizontal Equating

CLOSER LOOK

Equating and Vertical Scaling

COMMON QUESTION OR ISSUE

Q: I hear a lot about vertical linking, equating, and scaling. What do these terms mean?

EXPLANATION

First, consider the case of a state that develops three forms of a reading test for Grade 5:

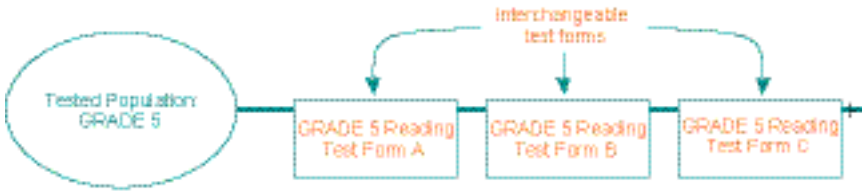


Figure 5.2: Horizontally equated test forms

Figure 5.2 is an example of *horizontal* equating. Test Forms A, B, and C may be administered in successive years. This is a common reason for linking and equating. Because all test forms in this example are linked and equated, scores can be used interchangeably from one year to the next.

On a conceptual level, *vertical* scaling is quite different from horizontal equating. Vertical scaling does not seek to produce interchangeable forms but instead seeks to place scores from tests at different grade *levels* on the same measurement scale so that the growth of individual students (or group of students) can be meaningfully tracked over time (Patz, 2007).

To further illustrate, consider the following:

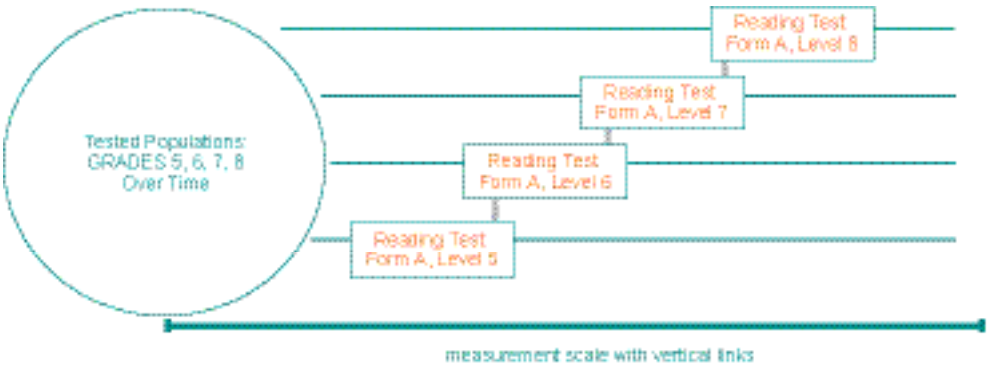


Figure 5.3: Vertically linked test forms (Form A in this example)

Figure 5.3 shows an example of vertical linking where different levels of Test Form A are linked across a common measurement scale. The vertical links are created to place scores from all levels of Test Form A along the same scale, allowing for comparisons over time.

Although many people use the term "vertical equating," this terminology is unacceptable because in these situations the assumptions for equating cannot generally be met; the process may be more aptly described as a method of linking test forms from one grade level to the next.

KEY IDEAS TO CONSIDER

Some educators might think that the reading curricula in grades 2 and 3 are similar enough that tests of these two levels could be *equated*. Certainly the grade 3 tests would have some material not found in grade 2 and vice-versa, but there could be enough overlap in the content domains between the two grades that a score on a grade 2 test form could have an equivalent (equated) score on the grade 3 form. Such a view would have to imagine that the grades 2 and 3 curricula are a single continuous curriculum with grade 2 material being relatively easier, grade 3 being more difficult material, and common material may be introduced in grade 2 and re-taught in grade 3. Similarly, grade 3 forms could be equated to grade 4, grade 4 to grade 5, etc. While it is conceptually possible to imagine a curriculum that spans several grades and a common scale across these grades, certainly there is a limit at some point to the meaning of a scale and the equivalence of scores across several grades.

From a technical point of view, the accumulation of sets of adjacent-grade-level tests would result in a vertical or cross-grade scale that could go from grade 3 to grade 8. Many such scales have been constructed by testing companies and state testing programs. However, a vertical scale for a state testing program presents some very challenging interpretation issues. At some point, the accumulation of adjacent-grade-level scales can stretch the meaning of the content and constructs being measured to a point where it is impossible to support some inferences. For example, it strains the imagination that a score on the grade 3 portion of the scale—the relatively easier material—has an equivalent score on the grade 8 portion of the scale. One can certainly *link* the different grade level scales together from grades 3 to 8, but across some grades the extension of the content/construct meaning cannot support the claim that tests at the different grades have the same substantive meaning. A vertically linked scale may be very useful in measuring progress over the grades, but the nature of what is being measured becomes very general and perhaps related more to underlying general ability rather than the achievement of specific grade-level content objectives.

It is also important to note that vertical linkages within an assessment program are also relatively difficult to accomplish. Vertical linking designs often require the use of common items at grades above and below the target grade level in order to establish links. Also, some assessment and subject matter experts suggest that significant shifts in content across grades can affect the meaning of the scores across the scale. Mathematics, science, and social studies content, for example, can shift considerably from one grade to the next. This practice raises the question as to how valid some vertical linkages might be for particular strands of content; it also calls into question how scores might be interpreted. If a fourth grader achieves a similar score on a level-4 science test compared to how some seventh graders do on their level-7 test, is it possible to say that the fourth grade student is doing science work at the seventh grade level?

Vertical linkages can be constructed vs. the *interpretations* of these vertical linkages. Vertical links, especially using adjacent grades, can be done very successfully and

have been found to be quite stable. However, interpretations drawn from links from adjacent grades are very different from interpretations of many of scores linked across two, three, or more grades.

Because of the potential for providing a method to track student progress over time, vertical linking and scaling is likely to continue to be a topic of interest. A psychometrician working on behalf of a statewide assessment program might recommend limiting the results of vertical linking to comparisons made at the *classroom level only* (not individual students) since group comparisons and scores are more reliable than individual results. However, this may change if the range of the scale is across a more limited set of grades. Vertical scale applications are likely to become more common as technical issues regarding their use are effectively resolved to the satisfaction of the educational measurement community.

KEY QUESTIONS TO ASK

Are the curriculum (content) standards vertically articulated?

Does our set of content standards permit a meaningful interpretation of a vertical scale?

Do we truly need a vertical scale, or might other statistical or analytical processes meet our needs?

What grade levels will be included in the vertical scale, and how much overlap might be present between grade levels?

How is growth defined? Do detailed specifications describing what will constitute "growth" exist?

What plans are in place for how the vertical scale will be developed and maintained?

Does an examination of the content expectations across grades suggest that the content goals/targets reflect a systematic progression in content over the grades?

How do we interpret vertical scale scores?

What do we do if a vertical scale shows that performances levels (e.g., cut scores) are not ordinal across grades?

5-G. Item Banking

Item Banking

COMMON QUESTION OR ISSUE

Q: What do we need to know about item banking? What role does it play in the equating process for our assessment program(s)?

EXPLANATION

One major reason to construct an item bank is to provide a repository of content for developing future test forms (and, later, equating test forms). The design of an item bank may occur early in the development of an assessment program and is often used to obtain items with IRT parameter values that can be used to construct test forms with prescribed IRT scale values before the forms are administered in an operational testing situation. Items being developed for a bank should be constructed to assess content standards for a particular grade, subject area or reporting category, and cognitive level. The item bank may also contain a variety of question formats including selected response, short answer, and extended response items. The items used on a test form are an approximate representation of the bank as a whole; that is, items in the bank reflect about the same proportion of item content and item types as specified for the operational test forms. Item banks are often over-built with the expectation that items will be lost due to public release or any other reason one might anticipate.

Items in an item bank have been field tested and may even have appeared on an operational test form. (Publicly released items are usually removed from the item bank as they are typically prohibited from use in future operational test forms.) IRT item parameters are estimated for each item after field testing or operational use, providing known item parameter values for all items in the bank. The parameters for these banked items define and fix the scale used in the assessment systems; this is sometimes described by saying "the bank defines the origin of the assessment scale."

The physical makeup of the item bank itself may vary from program to program, but computerized item banking systems are now the *de facto* standard among testing contractors. The item bank will often consist of the text, graphics, and a database of attributes associated with each item along with a record of classical and IRT statistical characteristics. It is important to clarify with test contractors not only the physical requirements for the item bank, but also **the plan or process of transferring the ownership, maintenance, and physical security of an item bank when moving from one test contractor to another.**

Item Bank Purpose and Uses for Equating

In some cases, the item bank is large enough to construct multiple test forms for use over the course of several years. In other cases, item banks are supplemented annually by incorporating field test items that have been embedded in operational

test forms created from the bank itself. Field test items are often brought onto the bank scale using the fixed parameter method described in this handbook under the heading "Equating with Concurrent or Simultaneous Calibrations," but other methods can be used as well.

Embedded field test items are generally not used in determining students' scores; from the test taker's perspective, these items often do not "count" even though students rarely know this. Rather, the field test items are scored and evaluated in terms of classical and IRT criteria, then assigned IRT parameter values by scaling them along with items from the bank (which have previous fixed values). Field test items that perform well are then added to the bank, thus replenishing the bank's item supply. As the number of items in the bank increases, developers are able to construct new test forms to match content specifications. Items for new test forms can also be selected to approximate predetermined psychometric values, such as average item difficulty or additional precision at one end of the ability scale.

Since items in a bank have IRT parameters calibrated to the desired measurement scale, a new operational test form can be developed to be a close match, statistically speaking, to an existing operational test form. As a result, scores from the new test forms are equated and can be used interchangeably with the scores of the older test. As time and resources allow, forms constructed from bank items are used as operational forms; after administration, the IRT values are re-estimated to check for stability.

Item bank values are frequently used when performance standards are set for an assessment program (Cizek & Sternberg, 2001; Cizek and Bunch, 2007). The most commonly used standard-setting procedures eventually relate the decision of a standard-setting panel (or panels) to a set of items. These items may constitute an intact operational form, or they may be banked items assembled specifically for the process of standard setting. In any case, the panel's decisions are used to locate points or cut scores on a scale, and the scale itself is defined by the bank values of the items used for the standard-setting procedure. Thus, performance standards or cut scores are set *in terms of the measurement scale defined by the items in the item bank*.

Item Bank Construction

The ideal item bank construction scenario involves collecting item data under testing conditions that are the same as, or a strong approximation of, the actual operational testing situation. In such cases, students' motivation and the activities of test administrators would closely resemble operational testing conditions. However, this circumstance is difficult to create: it requires "live" field testing of multiple test forms and sufficient time for post-test scoring, analyses, equating and perhaps even standard setting before results are reported. The pressure to return results quickly does not always allow for this kind of schedule.

In practice, developers use various combinations of procedures for constructing item banks. These procedures may use features of the "live" field test approach and features of the stand-alone approach. For example, after initial field testing, many states embed new field test items in subsequent operational test forms to build up the item bank. Ultimately, the approaches and procedures used to establish an item bank depend on two key factors: 1.) logistical flexibility and 2.) availability of resources.

Item Bank Development Designs

A typical approach to establishing an initial item bank uses an anchor item design. A set of common items are used along with unique items to create multiple field test forms. For example, consider a sixth grade reading test with 60 items:

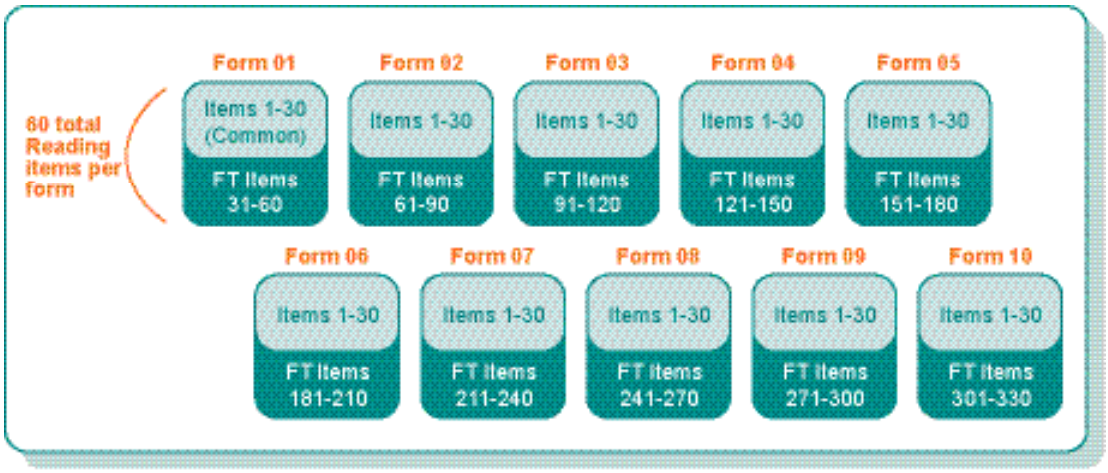


Figure 5.4: Using common items to build an item bank.
Unique field test (FT) items are administered along with common items (1-30) on all test forms.

Figure 5.4 shows 10 field test forms, with 30 items common to all 10 forms. For ease of illustration, the example shows the field test items assigned to the second half of the test. In practice, field test items would be interspersed in designated slots throughout the test form. The position of field test items should be recorded in the item bank, and efforts should be made to use the items in approximately the same positions when items are used on operational test forms.

Each form in Figure 5.4 would have 30 field test items unique to each specific form, resulting in a total field test of 330 Reading items. Reading and some other content areas have special considerations since subsets of items are associated with reading passages or other stimulus material. There is no requirement that a single anchor set be used to link all items to be banked. Pairs of forms, or sets of multiple forms, could be linked via different sets of common items.

KEY IDEAS TO CONSIDER

The main benefits of having an item bank are rooted in the flexibility it allows: the ability construct new test forms, pre-equate test forms, and assemble test content relatively quickly in response to a myriad of situations are major advantages.

However, the advantages of item banking also come at a cost. Advances in computer software have resulted in a proliferation of item banking systems that continue to evolve in terms of their complexity, sophistication, and expense. Evaluating these systems can be a difficult matter, because they are often proprietary systems. A key point to clarify is what the actual item bank consist of—is it text, images, item attributes, item statistics, database tables, item usage reports, etc., or some subset of these materials? In practice, the definition of "item bank" varies widely.

When item banks are used as part of an ongoing assessment program, IRT parameter values for item difficulty, discrimination, and the pseudo-guessing parameter are generally used as if they remain stable or constant when items from an item bank or another test form are used on a newly constructed test form. In some cases, however, the IRT values change or drift away from their bank values. Any substantial item parameter drift can compromise equating when IRT methods are employed. A variety of methods have been implemented for evaluating item drift including the Robust Z (Huynh & Rawls, 2007) and the .3 logit criterion (Miller, Rotou, & Twing, 2004).

KEY QUESTIONS TO ASK

Who owns the item bank content? Is any of the assessment content proprietary to the test contractor? Is this information well documented?

Who is responsible for maintaining the IRT parameter estimates stored in the item bank? How frequently are IRT parameters refreshed?

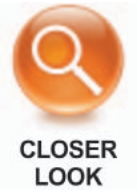
What processes are in place to check for item drift? How might it impact our assessment program?

How is the item bank *secured,* both physically and digitally?

Who *supports* the operation of the item bank system? What is the process for transferring custody of the item bank in the event that a different test contractor is selected to continue an existing program?

How easy/difficult is it to transfer the item bank from one test contractor's software system to a competing vendor's system? What can we expect such as transition to require in terms of our own time and effort?

5-H. Standard Setting and Accountability



Standard Setting and Accountability Issues

COMMON QUESTION OR ISSUE

Q: What role does equating play in standard setting and our overall accountability system?

EXPLANATION

Equating procedures can provide an assessment program with a way to develop new test content while still maintaining the meaning of the cut scores (performance standards) over time.

The practice of equating test forms also provides for a stable baseline to evaluate improvement, such as the percentage of the tested population achieving proficiency from year to year, even when items on the test forms are different.

KEY IDEAS TO CONSIDER

Equating plays an important role in the school accountability system because it allows scores to be compared from one year to the next. Thus, the technical quality of the equating methods used—and the documentation of the equating processes employed—are directly relevant to any accountability system intended to reflect annual growth and ongoing improvement.

A bank of items equated onto a common scale can be used to develop performance level descriptions for achievement levels defined by cut scores set on the scale.

In standard setting, panelists can respond to a test form constructed *specifically for standard setting* from a bank of calibrated items even if this test form has never been administered. Impact data on such a form can be projected from data collected from actual operational forms.

Every effort should be made in equating to maintain the cut scores as closely as possible. When vertical linking is employed, the same is true for the spacing and cross-grade ordering of the cut scores for various grades.

KEY QUESTIONS TO ASK

What does the accountability system *require*, directly or indirectly, in terms of producing and equating new test forms?

What impact does equating have on our established performance levels or cut scores?

Does the current or proposed standard setting process utilize actual test forms used by examinee populations, or will it be constructed from a bank of pre-calibrated items? How might this affect the written description of student ability at each performance level?

Technical Documentation

COMMON QUESTION OR ISSUE

Q: What type of documentation should contractors be required to produce when equating procedures are applied, or when an item bank has been developed and is being utilized?

EXPLANATION

Technical documentation is an essential part of all assessment procedures. All requests for proposals in contracts issued that include equating activities should include timely delivery of a concise but detailed equating technical report. This report should include, but not be limited to

- brief description of general assessment context
- description of the equating design and the equating procedure applied
- description of the sample used for equating comparing the demographics of the sample to the corresponding population characteristics
- results of the equating
- evaluation of the assumptions of the equating method used
- evaluation of the stability of the item parameters for anchor items
- analysis of any issues or problems encountered during equating and a description of how such issues were handled
- report of the raw score and scale score cut scores based on the equating compared to their corresponding cut scores on previous test forms
- report of the percentage of students in different performance levels based on the equating results compared to the corresponding proportions on previous test administrations
- recommendations for any modifications in the equating procedures

KEY IDEAS TO CONSIDER

The equating report should be written in sufficient detail such that the equating could be replicated by an independent contractor with the same results obtained. The integrity of the equating process is a critical aspect of the validity argument for the assessment and evidence supporting the validity of equating must be presented in the equating technical report. Well-documented, appropriate equating will satisfy a number of federal, and some state, reporting requirements.

KEY QUESTIONS TO ASK

Were there any problems or issues encountered during the assessment process?

Were there any steps taken that were *ad hoc* fixes to unanticipated problems?

How did the current equating compare to earlier equatings?

5-J. Inferences Based on Linking and Equating

Inferences Drawn from Equating

COMMON QUESTION OR ISSUE

Q: What inferences can be made as a result of the equating process?

EXPLANATION

The key inference that one hopes to support by equating is that students' performance on one test form has the same substantive meaning, and therefore the same meaning in terms of accountability decisions, as students' performances on the equated test form.

KEY IDEAS TO CONSIDER

The first step in supporting an inference of equivalence begins with test form construction. To be equated, test forms should be constructed according to the same test blueprints and specifications.

Practitioners should ask for the test blueprint and specifications and any other documentation that would support the inference that the forms are equated. Evidence from psychometric procedures should include data verifying that all assumptions made to support equating have been examined and reported. In common item equating, the evidence of the stability of the anchor items must be presented.

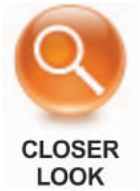
KEY QUESTIONS TO ASK

What test specifications are available to document the equating(s)? How closely were test specifications followed during the test construction process? How closely will they be followed for future forms?

Have any deviations from the test construction blueprints/specifications been taken into account during the equating process? If so, how are they documented, explained, and reported?

How strongly does the evidence gathered support the inference that equated test forms are equivalent and that test scores are interchangeable?

5-K. Quality Control Issues



Quality Control Issues

COMMON QUESTION OR ISSUE

Q: What controls will help ensure quality equating practices are used for our assessment programs?

EXPLANATION

Quality issues can take many forms. According to Kolen and Brennan, controlling for quality issues often takes more effort than other aspects of the equating process (2004, p. 306).

Generally, quality control issues can be segregated into problems that surface in the test design, development, and administration phases or the test scoring, analysis, and equating phases:

Test design, test development, and test administration problems include

- changes in test specifications
- item contexts that differ between forms and affect performance on anchor items
- anchor items that appear in very different locations among forms
- changes in anchor items
- misprints/errors
- keying problems
- cheating
- unintended accommodations (maps or periodic tables on walls, calculators, etc.)

Item scoring, analysis, and equating quality issues include

- non-standard scoring criteria or changes in scoring procedures
- not following proper/specified equating procedures
- problems with unreliable and/or inconsistent item performance or score distributions
- issues related to the processing of **conversion tables**
- item parameter drift

Data collection designs for equating require testing conditions to be as **standardized** as reasonably as possible for all test takers. A random groups design, for example, is viable only to the extent that Random Sample Group 1 takes Test Form A under the same conditions that Random Sample Group 2 experiences in taking Test Form B. If some students in Group 2 were given more time to complete Test Form B because of a power failure, fire drill, or any other reason, this could be considered an irregular testing condition.

In cases such as these, eliminating the non-standardized data could correct for the irregularity. However, the impact of such variations, and how best to deal with them, should be considered in the context of the overall assessment program (Kolen & Brannan, 2004, pp. 307-309).

KEY IDEAS TO CONSIDER

Standardization is a key aspect of test administration; standard testing conditions have a direct impact on equating quality. Allowing students more time than is specified, providing unauthorized accommodations, or not spiraling test forms randomly are examples of non-standardized conditions.

Sample size is also an important aspect of equating, because the relationship between sample size and the quality of the equating process is well established. Generally, larger samples are considered to result in better equating (Kolen, p. 288). Sample acquisition may be very carefully planned and executed in ways that are harmonious with the equating design, test specifications, and data collection model. Problems with administration, quality control, and other unplanned circumstances can strain the sampling for any given equating effort. Testing irregularities and administration problems, as discussed previously, can reduce the sample size near or below targeted levels and make any equating method more susceptible to error.

All equating procedures involve some amount of equating error, error in the estimation of item parameters, and the unreliability of the test in general. When multiple test forms are linked/equated, a "chain of equating" is formed. Because each step or link in the chain carries forth some amount of random or systematic error, the end links of the chain reflect a "build-up" of error. Thus, it is important to discuss with your contractor the ways in which error build-up might be monitored and controlled.

KEY QUESTIONS TO ASK

What specific quality control measures are in place to preserve the integrity of the equating process as much as possible?

What information does the contractor provide about equating error?

How will error build-up be controlled if we expect to link/equate multiple test forms over time?

Are the quality control measures documented? How are they checked/enforced?

Chapter Glossary:

Conversion Tables
Tables constructed in order to convert raw scores to scale scores and/or percentile ranks. Commonly used in pre-equating or other situations where rapid or immediate test results are desirable or required.

Rubric
A scoring guide for a performance task or a constructed-response item. Scoring rubrics contain a description of the requirements for varying degrees of success in responding to the question or performing the task.

Scaling
The process of associating numbers (or other ordered indicators) with the performance of individual test takers. Raw scores are transformed to *scale scores* using statistical methods. Typically, *scales* are constructed in ways that will help test users interpret the scores.

Standardization
The uniformity of test administration and scoring conditions from student to student and from place to place. Standardization helps make it possible to compare scores across situations. When tests are administered or scored in nonstandard ways, the results may not be reliably or validly compared to the test norms or performance criteria.

Test Blueprints
Written documents, often in chart form, that detail the number of questions to be included on a test, the item formats, and the content and skills that each set of items will assess. In the case of standards-based tests, it is important for the test blueprints to consider the performance standards as well as the content standards so that items cover the intended depth as well as breadth of the standards. In addition to guiding test development, test blueprints can be useful in preparing to take an examination.

References and Recommended Reading

Introduction and Chapter 1 References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Anderson, L., & Krathwohl, D. (Eds) (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* [Complete Edition]. New York: Longman Publishing Group.

Bloom, B.S., (1956). *Taxonomy of educational objectives.* New York, NY: Longmans, Green & Co.

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning.* New York, NY: McGraw-Hill Book Company

Brennan, R.L. (Ed) (2006). *Educational Measurement, 4th ed.* Westport, CT: Praeger Publishers

Feuer, M.J., Holland, P.W., Green, B.F., Bettenthal, M.W, & Hemphill, F.C. (Eds) (1999). *Uncommon Measures: Equivalence and linkage among educational tests.* Washington, DC: National Academy Press

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.).* New York, NY: Springer

Linking. Brennan, R.L. (Ed) (2006). *Educational Measurement, 4th ed*. Westport, CT: Praeger Publishers

Linn, R. (2008). *Validation of uses and interpretations of state assessments.* Washington, DC: Council of Chief State School Officers.

Marzano, R.J. & Kendall, J.S. (2007). *A new taxonomy of educational objectives (2nd ed.). Thousand Oaks, CA: Corwin Press.*

Redfield, D. (2001). Critical issues in large-scale assessment: A resource guide. Washington, DC: Council of Chief State School Officers.

Raw score. Parker, S.B. (2002). *McGraw-Hill dictionary of scientific and technical terms.* New York, NY: McGraw-Hill

Scaling. Kolen, M.J., & Brennan, R.L. (2004, pp. 52, 329). *Test equating, scaling, and linking: Methods and practices, 2nd ed. New York, NY: Springer*

Webb, Norman L., (1997). *Criteria for alignment of expectations and assessments in mathematics and science education.* National Institute for Science Education, University of Wisconsin-Madison; Washington, DC, the Council of Chief State School Officers.

Chapter 2 References

Allen, M.J., & Yen, W. M. (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland Press.

Baker, F.B., (2001). *The basics of item response theory, 2nd ed.* ERIC Clearinghouse on Assessment and Evaluation.

Crocker, L., & Algina, J. (1986) *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group

Downing, S., & Haladyna, T.M. (Eds). (2006) *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory.* Newberry Park, CA: Sage.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices, 2nd ed.* New York, NY: Springer

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Reliability. From Crocker, L., & Algina, J. (1986, p. 105) *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group; and Redfield, D. (2001). *Critical issues in large-scale assessment: A resource guide.* Washington, DC: Council of Chief State School Officers

Ryan, J., DeBiak, K. Osborn Popp, S., and Rivera, R. *An overview of item response theory,* presentation at the Arizona Educational Research Organization Conference, Arizona State University West, October 23, 2002.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, N.J: Lawrence Erlbaum Associates.

Chapter 3 References

Brennan, R.L. (Ed) (2006). *Educational Measurement, 4th ed.* Westport, CT: Praeger Publishers

Dings, J., Childs, R., & Kingston, N. (2002). *The effects of matrix sampling on student score comparability in constructed-response and multiple-choice assessments.* Washington, DC: Council of Chief State School Officers.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales.* Statistics for social and behavioral sciences. New York: Springer.

Downing, S., & Haladyna, T.M. (Eds). (2006) *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices, 2nd ed.* New York, NY: Springer

Patz, R. J., (2007). *Vertical scaling in standards-based educational assessment and accountability systems.* Washington, DC: Council of Chief State School Officers.

Popham, W. J. (1993). *Circumventing the high costs of authentic assessment.* Phi Delta Kappan, 7, 470-473.

Sinharay, S., & Holland, P. (2007). Is It Necessary to Make Anchor Tests Mini-Versions of the Tests Being Equated or Can Some Restrictions Be Relaxed? *Journal of Educational Measurement*, 44, 249-275.

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

Chapter 4 References

Allen, M.J., & Yen, W. M. (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland Press

Bejar, I. & Wingersky, M. (1982). A study of pre-equating based on item response theory. *Applied Psychological Measurement, Vol. 6, No. 3*, 309-325.

Baker, F.B., (2001). *The basics of item response theory, 2nd ed.* ERIC Clearinghouse on Assessment and Evaluation.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group

Cizek, G., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Newberry Park, CA: Sage

Cizek, G., & Sternberg, R. (Eds). (2001) *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum

Downing, S., & Haladyna, T.M. (Eds). (2006) *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory.* Newberry Park, CA: Sage

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices, 2nd ed.* New York, NY: Springer

Livingston, S. A. (2004). *Equating test scores (without IRT).* Retrieved September 12, 2007, from ETS web site http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

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

Mislevy, R.J. (1992) *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychgological Measurement,* 7, 201-210.

Taherbhai, H. & Young, M. (2004). Pre-equating: a simulation study based on a large scale assessment model. *Journal of applied measurement. Vol. 5, Part 3*, 301-318.

Chapter 5 References

Brennan, R.L. (Ed) (2006). *Educational Measurement, 4th ed.* Westport, CT: Praeger Publishers

Cizek, G., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Newberry Park, CA: Sage

Cizek, G., & Sternberg, R. (Eds). (2001) *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum

Cohen, J., Jiang, T., & Yu, P. (2008). *Information-weighted linking constants.* American Institutes for Research, Washington: DC. Unpublished paper.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales.* Statistics for social and behavioral sciences. New York: Springer.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory.* Newberry Park, CA: Sage.

Huynh, H., & Rawls, A., (2007). *A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model.* In Smith, E.V., & Smith, R.M. (Eds.), Rasch Measurement: Advanced and specialized applications. Maple Grove, MN: JAM Press

Ito, K., & Sykes, R. (1996) *A Comparison of Three Equating Approaches to A Random-Groups, Common-Forms Design.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices, 2nd ed.* New York, NY: Springer

Miller, G. E., Rotou, O., and Twing, J. S. (2004). Evaluation of the .3 logits screening criterion in common item equating. *Journal of applied measurement, 5(2),* 172-177.

Patz, R. J., (2007). *Vertical scaling in standards-based educational assessment and accountability systems. Washington, DC: Council of Chief State School Officers.*

CCSSO
Council of Chief State School Officers

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700 | Washington, DC 20001-1431
voice: 202.336.7000 | fax: 202.408.8072