Linking Adult Literacy Assessments

Robert J. Mislevy

Educational Testing Service

March, 1995

Contents

Linking Adult Literacy Assessments

Executive Summary

The National Literacy Act (NLA) of 1991 requires State Education Agencies to "gather and analyze data—including standardized test data—on the effectiveness of State-administered adult education programs, services, and activities, to determine the extent to which the State's adult education programs are achieving the goals in the state plan" [to enhance levels of adult literacy and improve the quality of adult education services]. These Federal evaluation requirements prompt interest in identifying standardized tests and methodologies that are appropriate for assessing the effectiveness of adult education programs, and for determining the feasibility of linking such tests in order to provide national trend data on program effectiveness.

Mislevy (1992) and Linn (1993) have laid out basic definitions, concepts, and approaches that have been developed over time to link educational tests for various purposes. This paper applies these general principles to the specific context of linking adult literacy tests. General principles are reviewed; basic linking approaches are described, with key ideas illustrated with examples from physical measurement and sports; and relevance to adult literacy assessment discussed. Conclusions about policy options appear below; the main body of the paper is intended to explicate their conceptual grounding.

Writing on the prospect of calibrating disparate tests to common national standards, Professor Andrew Porter (1991) wrote,

> If this practice of separate assessments continues, can the results be somehow equated so that results on one can also be stated in terms of results on the other? There are those who place great faith in the ability of statisticians to equate tests, but that faith is largely unjustified. Equating can be done only when tests measure the same thing. (p. 35)

Professor Porter's skepticism is justified. We are perhaps too familiar with correspondence tables that give exchangeable scores for alternate forms of standardized tests. But they work only because the alternate forms were constructed to meet the same tight specifications; equating studies and statistical formulas merely put into usable form the evidentiary relationships that were built into the tests.

Statistical procedures neither create nor determine relationships among test scores. Rather, the way that tests are constructed and administered, and the ways that the skills they tap relate to the people to whom they are administered, determine the nature of the potential relationships that exist in evidence that scores from the various tests convey. Statistical procedures, properly employed, can then be used to explicate these relationships and put them to use. Much progress has been made recently with statistical machinery for this purpose, but it is more complex than building simple once-and-for-all correspondence tables. These techniques are discussed in this review, *sans* equations, to give the reader an understanding of the kinds of linking procedures that are available, when they are appropriate, and approaches and caveats that apply when they are used.

The applicability of linking procedures to adult basic education (ABE) tests, with a special eye toward literacy tests, is driven by the facts that (i) adult education programs vary considerably with respect to the nature and level of skills they emphasize, and with respect to the kinds of students with whom they work; (ii) they use tests for a broad variety of diagnostic, instructional, and evaluative purposes; and (iii) these tests vary widely with respect to contexts, formats, and mixes of skills they tap. The first two conclusions are negative:

- *No single score can give a full picture of the range of skills that are important to all the different students in different ABE programs.*

- *No statistical machinery can translate the results of any two arbitrarily-selected ABE tests so that they provide interchangeable information about all relevant questions about student competencies and program effectiveness.*

What *is* possible? Three less ambitious, but more realistic, affirmative contingencies:

- *Comparing directly levels of performance across programs in terms of common indicators of performance on a market basket of consensually-defined tasks in standard conditions.* Some aspects of competence, and assessment contexts for gathering evidence about them, will be considered useful by a wide range of programs, and components of an assessment system can solicit information in about them in much the same way for all. However, these "universal" assessments—and in particular pre-post comparisons with such assessments—would provide seriously incomplete information to evaluate the effectiveness of programs, to the extent that their focus did not match program objectives.

- *Estimating levels of performance of groups or individuals within clusters of programs with similar objectives—possibly in quite different ways in different clusters—at the levels of accuracy demanded by purposes within clusters, with common assessments focused on those objectives.* These components of programs' assessments might gather evidence for different purposes, types of students, or levels of proficiency, to complement information gathered by "universal" components.

- *Making projections about how students from one program might have performed on the assessment of another.* When students can be administered portions of different clusters' assessments under conditions similar to those in which they are used operationally, we can estimate the joint distribution of results on those assessments. These studies are restricted as to time, place, program, and population, however. The more the assessments differ as to their form, content, and context, the more uncertainty is associated with these projections; the more they can be expected to vary with students' background and educational characteristics; the more they can shift over time; and the more comparisons of program effects become untrustworthy.

Linking Adult Literacy Assessments

Background

As defined by the National Literacy Act (NLA) of 1991, literacy involves "an individual's ability to read, write, and speak in English, compute, and solve problems at levels of proficiency necessary to function on the job and in society, to achieve one's goals, and to develop one's knowledge and potential." The NLA requires State Education Agencies to "gather and analyze data—including standardized test data—on the effectiveness of State-administered adult education programs, services, and activities, to determine the extent to which the State's adult education programs are achieving the goals in the state plan" [to enhance levels of adult literacy and improve the quality of adult education services] (*Federal Register,* 1992). These Federal evaluation requirements have prompted interest in identifying standardized tests and methodologies that are appropriate for assessing the effectiveness of adult education programs, and for determining the feasibility of linking such tests in order to provide national trend data on program effectiveness (Pelavin Associates, 1994a, p. 1).

Basic definitions, concepts, and approaches that have been developed over time to link various tests for various purposes are presented in Linn (1993) and Mislevy (1992). The present paper brings these ideas to bear on the specific problem of linking adult literacy tests. General principles are briefly reviewed; basic linking approaches are described, with key ideas illustrated with examples first from athletics then from literacy assessment; and finally, conclusions with implications for policy options are drawn.

A quick overview of adult literacy tests

The diversity of both the objectives and the participants served by adult education programs reflects a broad and multidimensional definition of literacy. Accordingly, adult education programs vary considerably with respect to the nature and level of skills they emphasize, and with respect to the kinds of students with whom they work. Moreover, they use tests for a broad variety of diagnostic, instructional, and evaluative purposes. Both the nature of the instruction and the purpose of testing determine the kinds of tests will be appropriate. Development Associates' (1992) recent survey of adult education tests found the following to be most widely used:

- **Tests of Adult Basic Education (TABE)** (used by 67.7% of the ABE programs, as reported in Development Associates, 1992, national evaluation) measures academic skills in adult education classes, especially reading comprehension, mathematics concepts and applications, mathematics computation, language mechanics, language expression, reading vocabulary and spelling.

- **Slosson Oral Reading Test (SORT)** (used by 22.9% of the ABE programs) measures oral word recognition. Originally designed for elementary and secondary school students, the test also is being used for adults.

- **Adult Basic Learning Examination (ABLE)** (used by 21.0% of the ABE programs) measures a mix of academic and some life skills.

- **Wide Range Achievement Test-Revised (WRAT-R)** (used by 20.1% of the ABE programs) measures three basic academic coding skills of reading, spelling, and arithmetic.

- **Comprehensive Adult Education Assessment System (CASAS)** (used by 14.2% of the ABE programs) includes a series of assessments, such as the Life Skills Appraisal, Life Skills Survey Achievement Tests, and Life Skills Certification Tests.

Two additional tests are also relevant to question of linking adult education tests, the first because of its use for predicting scores on the General Educational Development (GED) test and the second because of its ties to recent national surveys of adult literacy:

- **Official General Educational Development Practice Tests (OGEDPT)** measure academic skills in writing, social studies, science, mathematics, and interpreting literature and the arts. These tests are designed for adults preparing to take the full-length GED.

- **Tests of Applied Literacy Skills (TALS)** measure functional prose, document, and quantitative literacy within many life skill items.

At first blush, it might seem most convenient to simply build correspondence tables for all these tests (or at least for subtests ostensibly "measuring the same thing"), whereby the user could take a score from any test, find the right row and column in a table, and obtain the equivalent score on any of the other tests. This is basically how users of standardized achievement tests have, for years, obtained equivalent scale scores

and percentiles from different forms of a given test form; the formulas and the estimation procedures must be thoroughly worked out by now. The following discussion makes it clear, however, that things are not so simple. Only in rare cases is it possible to fully characterize in a correspondence table the evidential value of Test Y scores for all the inferences one would have used Test X scores for—and in these cases it is possible only because the tests were constructed from the same test specifications, or "blueprint." It is possible nevertheless to characterize and exploit the evidential value of Test X and Test Y scores more broadly—though resorting to more complex methodologies, suffering information loss to varying degrees, and facing the risk of certain systematic errors in predictions and estimation of effects. These methodologies, risks, and indications of their relevance to adult education assessments, are addressed in the sequel.

A hypothetical analogy from athletics

As defined by the National Fitness Act (NFA) of 1991, physical fitness involves "an individual's ability to carry out physical activity and participate in team or individual athletics, in recreation or competition, at levels of proficiency necessary to function in athletic activities and everyday life, to achieve one's goals, and to develop one's capabilities and potential." The NFA requires State Education Agencies to "gather and analyze data—including standardized test data—on the effectiveness of State-administered adult physical fitness programs, services, and activities, to determine the extent to which the State's adult fitness programs are achieving the goals in the state plan" [to enhance levels of adult fitness and improve the quality of adult fitness services] (*Federal Register,* 1992). These Federal evaluation requirements have prompted interest in identifying standardized tests and methodologies that are appropriate for assessing the effectiveness of adult fitness and athletic training programs, and for determining the feasibility of linking such tests in order to provide national trend data on program effectiveness.

The diversity of both the objectives and the participants served by adult fitness programs reflects a broad and multidimensional definition of fitness. Accordingly, adult fitness programs vary considerably with respect to the nature and level of skills they emphasize, and with respect to the kinds of students with whom they work. Some programs focus on increasing the strength and stamina of senior citizens; others focus on special problems such as weight loss or rehabilitation from injury; still others increase fitness through instruction and participation in particular sports, as in track and field, basketball, and swimming programs. These programs use tests for a broad variety of diagnostic, instructional, and evaluative purposes. Both the nature of the instruction and the purpose of testing determine what kinds of tests will be appropriate.

Is it possible to somehow "link" all the various tests these fitness programs use? No. There can be no direct one-to-one correspondence between, say, the improved times of young adults marathon runs in one program with the lengthening walker-assisted ventures of surgery-recovery patients in another program, or the increasing proportions of free-throws made by participants in the basketball clinic. Programs encounter an inevitable tradeoff: The better a test focuses specifically on information tailored to its objectives and participants, the less likely it is to overlap the information provided by tests tailored to other programs. Yet such tests are invaluable in evaluating its success.

These specifically-useful indicators can be supplemented nevertheless by broader measures of fitness, still of more or less value in evaluating a program's specific objectives, that can be employed to provide some useful information across broader ranges of programs (though still probably not all). Examples might include treadmill-based measures of heart and pulmonary function, which could, with effort, be standardized and compared across programs. These measures would monitor aspects of fitness presumably valued in all programs, but, since they are more or less central to the objectives of different programs, could not serve as the sole basis of program evaluations. Nor could run-times, free-throw percentages, or batting averages be "calibrated" to heart and pulmonary measures, any more than these specifically useful indicators could be "calibrated" to one another (although one could probably determine approximate equivalents between, say, 10K and marathon times, or pole-vaulting heights with bamboo and fiberglass poles). The diversity of objectives, consistent with the breadth of the NFA, necessitates that thorough evaluations of individual programs would need to consider both broadly common indices, and indices tailored and validated for their more varied specific situations.

Necessary concepts from educational assessment

The terms "assessment" and "test" refer in this paper to systematic ways of gathering and summarizing evidence about student competencies. Assessments provide data such as written essays, correct and incorrect marks on an answer sheet, and students' explanations of the rationales for their problem solutions. This *data* becomes *evidence*, however, only with respect to inferences about students' competence—perhaps concerning individual students, the group as a whole or particular subgroups, or even predictions about future performance or outcomes of instruction. Our purposes for assessing and our conception of the nature of competence drive the form an assessment takes. An assessment that produces solid evidence at reasonable costs for one mission can prove unreliable, exorbitant, or irrelevant for another.

Suppose that Assessment X provides evidence about instructional or policy questions involving student competencies. In particular, we might want to know whether individuals can perform at particular levels of competence—i.e., whether they are achieving specified educational standards. Suppose that with appropriate statistical tools, we could provide answers, each properly qualified by an indication of the strength of evidence. Let's also suppose someone wants to "link Assessment Y to Assessment X." This notion stems from a desire to address these same questions, posed in terms of Assessment X, when what we observe is students' performances on Assessment Y.

As noted above, adult education programs use tests for a wide variety of purposes. What's important is that different kinds and amounts of evidence must be gathered to suit different purposes. Certain distinctions among purposes are pertinent in this regard to our discussion about linking:

- Will important decisions be based on the results; that is, is it a "high stakes" assessment? A quiz to help determine what a student should work on today is a low-stakes assessment; a poor choice is easily remedied. An test to determine whether they should receive a high-school equivalency certificate is high-stakes at the level of the individual. An assessment to evaluate programs for continued funding is high-stakes at the level of programs. An assessment that supports decisions of consequence must provide commensurately dependable evidence, just as criminal conviction demands proof "beyond a reasonable doubt."

- Do inferences concern a student's comparative standing in a group of students—that is, are they "norm-referenced"—or do they gauge the student's competencies in terms of particular levels of skills or performance—that is, are they "criterion referenced?"

- Do inferences concern the competencies of individual students, as with the Tests of Adult Literacy Skills (TALS), or the distributions of competencies in groups of students, as with the Survey of Young Adult Literacy from which TALS originated? When the focus is on the individual, one must gather enough evidence about each student to support inference about him or her specifically. On the other hand, a bit of information about each of several students—too little to say much about any of them as individuals—can suffice in the aggregate to monitor of performance in a program or a state.

Assessments as operational definitions

Recall the NLA's statement that literacy involves "an individual's ability to read, write, and speak in english, compute, and solve problems at levels of proficiency necessary to function on the job and in society, to achieve one's goals, and to develop one's knowledge and potential." This is all well and good in the abstract, but it is an actual specific assessment that the student ultimately encounters. An actual test must be constructed from a set of test specifications, which amounts to a "blueprint" of what a particular assessment should comprise: the kinds and numbers of tasks it will comprise, the way it will be carried out, the processes by which observations will be summarized and reported. This level of specification determines an *operational definition* of competence. Quality-control statistician W. Edwards Deming describes how similar processes are routinely required in industry, law, and medicine:

> Does pollution mean, for example, carbon monoxide in sufficient concentration to cause sickness in 3 breaths, or does one mean carbon monoxide in sufficient concentration to cause sickness when breathed continuously over a period of 5 days? In either case, how is the effect going to be recognized? By what procedure is the presence of carbon monoxide to be detected? What is the diagnosis or criterion for poisoning? Men? Animals? If men, how will they be selected? How many? How many in the sample must satisfy the criteria for poisoning from carbon monoxide in order that we may declare the air to be unsafe for a few breaths, or for a steady diet?
>
> Operational definitions are necessary for economy and reliability. Without an operational definition, unemployment, pollution, safety of goods and of apparatus, effectiveness (as of a drug), side-effects, duration of dosage before side-effects become apparent (as examples), have no meaning unless defined in statistical terms. Without an operational definition, investigations on a problem will be costly and ineffective, almost certain to lead to endless bickering and controversy. (deming, 1980, p. 259)

A study of pollution in cities might include several of these operational definitions, to provide a fuller picture of their environments. No single operational definition can tell "the full story"; no single indicator can capture "the truth." For this reason, educational assessments are often comprised of multiple scores for each student, to provide a fuller picture of their competencies, and evaluations of educational programs often use multiple assessments. Operational definitions of aspects of literacy that all fall under the broad umbrella of the NLA definition include the sort with its focus on oral

word recognition, the WRAT with an emphasis on basic decoding and computational skills (intentionally designed to "eliminate, as totally as possible, the effects of comprehension;" Jastak & Wilkinson, 1984, p.1), and the TALS with its accent on reasoning in real-life contexts.

Test theory

The most familiar tools we have for linking assessments evolved under the paradigm of "mental measurement," spawned a century ago in an attempt to "measure intelligence." The measurement paradigm posits that important aspects of students' knowledge or skills can sometimes be represented by numbers that locate them along continua, much as their heights and weights measure some of their physical characteristics. From a behavioral point of view, the variable of interest might be the proportion of correct answers a student would give on every item in a large domain. From a cognitive point of view, the variable might be a level on a developmental scale such as the American Council on the Training of Foreign Languages (ACTFL) Reading guidelines, excerpts from which appear as Table 1. Test theory views observed scores as noisy manifestations of these inherently unobservable variables, and attacks the problem of inference in the face of measurement error. We discuss necessary elements of classical test theory in the section about "equating," and of item response theory in the section on "calibration."

[[Table 1: ACTFL reading guidelines]]

We do well to remember that "traits" achievement tests purportedly "measure," such as "mathematical ability," "reading level," or "physics achievement," do not exist *per se*. Contemporary conceptions of learning do not describe developing competence in terms of "increasing trait values," but in terms of constructing and reconstructing mental structures that organize facts and skills ("schemas"); of learning how to plan, monitor, and, when necessary, switch, problem-solving strategies ("metacognitive skills"); and of practicing procedures to the point that they no longer demand high levels of attention ("automaticity"). Test scores tell us something about what students know and can do, but any assessment setting stimulates a unique constellation of knowledge, skill, strategies, and motivation within each examinee, and different settings stimulate different unique constellations. The mental measurement paradigm is useful to the extent that the patterns of behavior it captures guide instructional or policy decisions in ways consistent with our

conceptions of the acquisition of competence. Educational psychologists Richard Snow and David Lohman put it like this:

> Summary test scores, and factors based on them, have often been thought of as "signs" indicating the presence of underlying, latent traits. … An alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the similarity or overlap of this sampling, is equally justifiable and could be theoretically more useful. The evidence from cognitive psychology suggests that test performances are comprised of complex assemblies of component information-processing actions that are adapted to task requirements during performance. The implication is that sign-trait interpretations of test scores and their intercorrelations are superficial summaries at best. At worst, they have misled scientists, and the public, into thinking of fundamental, fixed entities, measured in amounts. Whatever their practical value as summaries, for selection, classification, certification, or program evaluation, the cognitive psychological view is that such interpretations no longer suffice as scientific explanations of aptitude and achievement constructs. (Snow & Lohman, 1989, p. 317)

We may therefore build assessments around tasks suggested by the appropriate psychology, but summarize evidence using measurement models. In some applications, we may wish to model competencies with detailed structures suggested by the psychology of learning in the domain. An example is tutoring an individual students' foreign language proficiencies; the ACTFL scale doesn't capture the distinctions we'd need to help a Mid-Novice become a High-Novice. In other applications, such as tracking a students' progress in broad strokes, the ACTFL scale may suffice. Furthermore, once we recognize that measurement model results are at best gross summaries of aspects of students' thinking and problem-solving, we are obliged to identify the contexts that circumscribe their usefulness. This perspective underlies the concern for verifying the usefulness of, say, reading tests developed for U.S. middle-school students with adult ESL students.

The key question from a statistician's point of view is "How probable is this particular observation, from each of the possible 'true' values"? The answer—the so-called "likelihood function" induced by the response—embodies the information that the observation conveys about competence, in the way competence is being operationally defined. If the observation is equally likely from students at all values of the variables in the competence model, it carries no information. If it is likely at some values but not

others, it sways our belief in those directions, with strength in proportion to how much more likely the observation is at those values. Terms such as "reliability" and "accuracy" can be thought of as the sharpness or diffuseness of the likelihood function that an observed score induces for inference about true score.

A survey of methods for linking tests

Overview

The central problems of "linking assessments" are (1) explicating the relationships among the evidence two assessments provide about conjectures of interest, and (2) figuring out how to interpret this evidence correctly. Summaries of methods that have been devised to tackle this problem in educational assessment appear below. Table 2 summarizes their key features. Following sections describe and illustrate each method in greater detail. An integrative discussion and policy implication appear later in the paper.

[[Table 2: Methods of linking]]

- *Equating.* Linking is strong and straightforward *IF* Assessment Y has been constructed from the same blueprint as Assessment X. Under these carefully controlled circumstances, the weight and nature of the evidence that the two assessments provide about a broad array of questions is practically identical. By matching up score distributions from samples of similar students, we can construct a one-to-one table of correspondence between scores on X and scores on Y with the property that any question that could be addressed using X scores can be addressed in exactly the same way with the same accuracy with transformed Y scores, and vice versa.

- *Calibration*. A different kind of linking is possible if Assessment Y has been constructed to provide evidence about the same conception of competence as Assessment X, but with different amounts or different kinds of evidence. The term *calibrating* applies, in analogy to physical measurement: Scales are adjusted so that the expected score of a student is the same on all tests that are appropriate to administer to him or her.[1] Unlike equating, where the focus is matching tests to one another directly, calibration

[1] Some writers use the terms "equating" and "calibrating" interchangably to describe what I call calibrating. There are enough differences in procedures and properties to maintain the distinction.

draws results of different assessments to a common frame of reference, and thus to one another only indirectly.

Some properties of calibration are disconcerting to those familiar with only equating: As a consequence of different weights of evidence in X and Y data, the procedures needed to give the right answers to some "X questions" from Y data give the wrong answers to others. It is possible to answer X questions with Y data if a calibration model holds, but generally *not* by means of a single correspondence table. We discuss three settings in which calibration applies: (1) constructing tests from essentially the same blueprint, but with differing lengths; (2) using item response theory to link connect responses to a collection of items to the same construct; and (3) soliciting judgments in terms of abstractly defined criteria.

• *Projection.* If assessments are constructed around different types of tasks, administered under different conditions, or used for purposes that bear different implications for students' affect and motivation, then mechanically applying equating or calibration formulas can prove seriously misleading; X and Y do not "measure the same thing." This is not merely a matter of stronger or weaker information, but of qualitatively different information. If it is sensible to administer both X and Y to any student, statistical machinery exists to (1) estimate, in a linking study, relationships among scores from X and Y, and other variables in a population of interest, then (2) derive *projections* from Y data about what the answers to the X questions might have been, in terms of a probability distribution for our expectations about the possible outcomes.

As X and Y become increasingly discrepant—as when they are meant to provide evidence for increasingly different conceptions of competence—the evidential value of Y data for X questions drops (there is more uncertainty associated with the projection), and projections become increasingly sensitive to other sources of information. Importantly, the relationship between X and Y can differ among different groups of students, and can change over time in response to policy and instruction.

• *Moderation*. Certain assessment systems obtain Test X scores from some students and Test Y scores from others, under circumstances in which it isn't sensible to administer both tests to any student; literature students take a literature test, for example, and history students take a history test. There is no pretense that the two tests "measure the same thing," but scores that are in some sense "comparable" are desired nevertheless.

Whereas projection evaluates the evidence that results on one assessment provide about likely outcomes on another, *moderation* simply aligns scores from the two as to some measure of "comparable worth." The way that "comparable worth" is determined distinguishes two varieties of moderation:

Statistical moderation matches up X and Y score distributions, sometimes as a function of joint score distributions on a third "moderator test" that all students take. That is, a score on X and a score on Y are deemed comparable if the same proportion of students in a designated reference population (real or hypothetical) attains scores at or above that level. The score levels that end up matched can depend materially on the choice of the reference population and, if there is one, the moderator test. Historically, these procedures have been discussed as a form of "scaling" (Angoff, 1984).

Social moderation uses judgment to match levels of performance on different assessments directly to one another (Mislevy, 1992; Wilson, 1992). This contrasts with the judgmental linking discussed under calibration, which maps performances from different assessments to a common more abstractly defined variable, and under projection, which evaluates the evidence that judgmental ratings obtained in one context hold for another context.

Equating, calibration, and moderation address the correspondence between single scores from two assessments. Two assessments might each have multiple scores, however, and these approaches could apply to matched pairs of scores. Projection can address joint relationships among all scores from multiple assessments simultaneously.

Equating

Basic concepts of equating

Selection and placement testing programs update their tests periodically, as the content of specific items becomes either obsolete or familiar to prospective examinees. New test forms are constructed according to the same blueprint as previous forms: the same number of items, asking similar questions about similar topics in the same ways. SAT Mathematics test developers, for example, maintain the balance among items with no diagrams, with diagrams drawn to scale, and diagrams not drawn to scale, along with a hundred other formal and informal constraints. Pretest samples of examinee responses to the new items are gathered along with responses to items from previous forms, and items'

statistical properties from pretest samples may additionally be used to select items for new forms. The objective: "parallel" test forms, which provide approximately equivalent evidence for a broad range of potential conjectures. *Equating* makes slight adjustments in the results of such test forms (which were expressly constructed to make such adjustments negligible!) by aligning the distributions of scores from the same or similar students on the two forms. In technical terms, equated scores from two tests give rise to essentially identical likelihood functions for inferences about competence.

Equating in physical measurement

Two hundred years ago, Karl Friedrich Gauss studied how to estimate the "true position" of a star from multiple observations—all similar, but not identical because of the inherent imperfections of the telescope-and-observer system. The nature of the measuring instrument determines the typical size and distribution of the measurement errors. If two instruments react to exactly the same physical property with exactly the same sensitivity, their readings can be *equated,* in the following sense: A table of correspondence can be constructed so that a value from one instrument has exactly the same interpretation and measurement error distribution as the same value would from the other instrument.

For equating to occur, it is not necessary that either measure be especially accurate; indeed we expect that there will be some difference between, say, "true temperature" and "observed temperature readings," even after the different thermometers have been equated. The observed reading is always the true temperature perturbed by some "noise." What is essential for strict equating to hold, however, is that the distribution of noise is the same for all the thermometers to be equated.

If the noise is truly random and independent of the true temperature, the average for a number of, say, patients in a hospital is about the same as the true average. On the whole, though, the thermometer readings are spread out a little more than the true values, because the measurement errors add variance to the collection of measures. (We shall see how this point becomes important in *calibration*.)

Parallel time charts in horse-racing

Handicappers of thoroughbred races know that some tracks are "faster" than others at any given distance, so that horses' times at different tracks are not directly comparable. As an example, the

chart shown as Table 3 (extracted from Davidowitz, 1977, p. 204) maps times from $1^1/_8$ mile races at Belmont and Aqueduct onto a common "speed figure" metric. Such a chart is constructed by researching winning times for each racing class at the tracks in question. Aqueduct is a slower track for $1^1/_8$ mile races because a horse must round two turns rather than one as at Belmont. The performance required is virtually identical, the distances are the same, and the time differential is slight--only about 2 seconds in a races that takes nearly minutes. These differentials apply to older and younger horses and at all seasons, even though horses' ages and time-of-year affect typical winning times. This relationship is tantamount to an equating.

[[Table 3: race-track chart]]

Equating in educational assessment

Edgeworth (1888, 1892) and Spearman (1904, 1907) launched classical test theory (CTT) around the turn of the century by applying the ideas of true-score measurement to tests. CTT views the average (or, equivalently, the total) of 1-for-right/0-for-wrong results from numerous test items as an imperfect measure of an examinee's "true score." While each of the separate items taps specific skills and knowledge, a score from such a test captures a broad general tendency to get items correct, and provides information for conjectures about competencies also defined broadly (Green, 1978). Different tests drawn from the same domain correspond to repeated measures of the same true score.

The indicator of a test's accuracy under classical test theory is *reliability*, a number between 0 to 1 gauging the extent of agreement between different forms of the test. Reliability is approximated by the correlation between two parallel test forms; it is also theoretically equivalent to the square root of the correlation between observed scores on one of these forms and the "true scores" that they estimate imperfectly. This definition reflects the classic norm-referenced usage of tests: lining up examinees along a single dimension for selection and placement. A test that measures examinees' true scores perfectly accurately is useless for this purpose if all their true scores are the same, so this test has a reliability of 0 for this population—although it might separate examinees in a different population quite well, and have a reliability closer to one.

A high reliability coefficient is important when scores are used for high-stakes norm-referenced uses with individual students, because a different sample of items of the same kind would lead to the same decision about most of them. This same index can be

less important, even inappropriate, for other purposes. As a first example, a shorter and less reliable test can suffice for less consequential norm-referenced decisions—a diagnostic pretest to determine whether to focus on certain literacy skills. Second, when the purposes of assessment are criterion-referenced, evidence about the accuracy with which an individual's competencies are gauged is more important than evidence about how he or she compares with other students. Third, accurate estimates of aspects of the distribution of true scores in a group of students can be obtained from short tests, even if measurement for individuals is quite inaccurate (Lord, 1962). The National Assessment for Educational Progress (NAEP) and the Survey of Young Adult Literacy (SYAL), designed to provide information at the level of groups rather than individuals, exploit this result by giving each student a small sample of items from a large pool, and amassing substantial evidence about groups on a broader range of tasks than could be administered to any single student (Mislevy *et al.,* 1992).

Equating applies to tests that embody not merely the same general statement of competence, but structurally equivalent operational definitions. But even with the care taken to create parallel test forms, scores may tend to be a bit higher or lower on the average with the new form than with the previous one. Its scores may spread examinees out a little more or less. An equating procedure is meant to adjust for these overall differences, so that knowing which particular form of a test an examinee takes no longer conveys information about the score we'd expect. There are a variety of equating schemes (see Petersen, Kolen, & Hoover, 1989), but the "equivalent groups" design captures the essence:

1. Administer Test X and Test Y to randomly-selected samples of students from the same specified group of students—typically, a group representative of students with whom the tests will be used.

2. Make a correspondence table matching up the score on Test X that 99% of the sample was above, with the score on Test Y that 99% of the sample was above. Do the same with the 98% score level, the 97% level, and so on. It is usually necessary to "smooth" the relationships, if for no other reason than test scores are integers rather than continuous numbers. If the X and Y distributions have similar shapes, we may be able to align the distributions sufficiently well by just matching up the means and standard deviations of the X and Y samples; this is "linear equating."

3. Once a correspondence table has been drawn up, look up any score on Test Y, transform it to the tabled Test X score, and use the result exactly as if were an observed Test X score with that value. The same table can be used in the same way to go from X scores to equated Y scores.

Long tests constructed from a given domain of tasks correspond to more accurate measures than short tests drawn from the same domain. Equating, strictly defined, applies to two similarly constructed long tests, or two similarly constructed short ones—but not a long and a short test, even if they are built around the same skills. The calibration section below discusses the problems that arise if we use equating procedures to link a long test and a short one.

Some connections with adult literacy tests

Several commonly used adult education tests are available in parallel forms that provide equated scores. The OGEDPT offers four parallel "half length" forms, which are *equated* to one another but are only half as long as the full GED they are meant to predict. Although OGEDPT half-length forms have the same kinds of items as the GED and their scores on are reported on the GED standard scales, they are *not* equated to the GED; rather, they are *calibrated* to the GED scale, as described in the Calibration section below.

TABE too has parallel forms that equated to each other. TABE can also be used to predict GED scores, but because these tests tap somewhat different mixes of skills, this relationship is neither an equating nor a calibration, but an prediction based on *projection* (see the Projection section below). TABE predictions of GED scores could tend to overshoot or undershoot actual GED performance in distinguishable kinds of examinees.

What happens when the formulas of equating are applied with scores on tests not constructed to be parallel? By the way the resulting correspondence tables are constructed, there are *some* inferences that matching scores give the same correct answer to; for example, for this group of examinees, at this point in time, what proportion have scores above *Z*? But these same tables can give seriously misleading and nonequivalent answers to different questions, as the examinees in question differ in time and experience from those from whom the table was constructed. These points will be explored further in the sequel, especially in the section on Statistical Moderation.

Comments on equating

The test construction procedures and the statistical procedures of equating constitute an inseparable package. When they are applied in concert, equated scores from parallel test forms provide virtually exchangeable evidence about students' behavior on the same general domain of tasks, under the same specified standard conditions. Aside from sampling error, the same equating function would be obtained from examinees of different instructional or demographic backgrounds. When equating does work, it works because of the way the *tests* are constructed, not simply because of the way the linking data are collected or the correspondence tables are built. Tests constructed in this way can be equated whether or not the tests are actually "measuring something" in the deeper senses discussed in the following section—although of course determining what skills and knowledge their scores reflect is crucial for justifying their use.

Calibration

Basic concepts of calibration

"Calibration" relates observed performance on different assessments to a common frame of reference. Properly calibrated instruments have the same expected value for measuring an object with a given true value. These instruments provide evidence about the same underlying variable, but, in contrast to equating, possibly in different amounts or by different methods. The focus is on inference from the measuring instrument to the underlying variable, so calibrated instruments are related to one another only indirectly. This too contrasts with equating, where the focus is on matching up observed performance levels on tests to one another directly. After illustrating properties of calibration in the context of temperature, we discuss three cases of calibration in educational assessment.

In physical measurement, scientists "calibrate" measuring instruments that differ in their accuracy, or are accurate in different ranges of the quality they are designed to measure. As an example, the warmer it gets, the faster the snowy tree cricket (*Oecanthus fultoni*) chirps. The relationship between temperature and cricket chirps isn't as strong as the relationship between temperature and the density of mercury, and it isn't very useful below freezing or above about 100°. "Calibrating" a cricket means transforming the counts of chirps so that for any true temperature, the resulting estimate of temperature is

as likely to be above the true value as below it. The *Encyclopedia Brittanica* gives "the number of chirps in 15 seconds + 40" as an approximation for Fahrenheit temperature.

If the true temperature is 65°, five calibrated temperature readings from a cricket's chirps could be 57°, 61°, 66°, 71°, and 75°—averaging around the true value, but quite spread out. In contrast, five readings from a backyard thermometer, say Thermometer X, could give readings of 63°, 64°, 65°, 65°, and 66°—also averaging around the true value, but with much less spread. If we have a reading of 65° from Thermometer X or from a cricket, our best estimate of the true temperature is 65° in either case—but the evidence that the true temperature is around 65° rather than 60° or 70° is much stronger with Thermometer X than with the cricket.

Different tests that are calibrated to the same scale "measure the same thing" but may have different amounts or distributions of "measurement error." In technical terms, they give rise to likelihood functions that have their highest value at the same point along the scale, but can differ as to shape and dispersion. Consequently, scores from different tests that are calibrated in one of the ways discussed below possess a seemingly paradoxical pair of properties:

1. *The observed score is an unbiased estimate of the "true score" of an individual examinee, and the observed mean of a group of examinees is an unbiased estimate of the mean of their true scores.*

It would thus seem that a table of corresponding scores between the two tests would suffice for translating information from one test into terms of information from the other test. This is true for inferences aimed at point estimates of these means, and it works because the test scores can be aligned so that best single-number estimates match up. But even after this is done, the nature and amount of uncertainty associated with those estimates generally does not match up, and this distorts other inferences one might wish to make. The shape and the spread of the distribution of scores from calibrated tests depends on the amount of measurement error associated with the scores from the two tests. In particular,

2. *The distributions of a group of examinees' scores on correctly calibrated tests, while agreeing as to average, may differ substantially as to spread of scores and proportions of examinees above cut-points.* Distributions of observed scores from less reliable tests are more spread out than distributions of observed scores

of more reliable tests, even if the two tests are properly calibrated to the same underlying variable.

One could build a different correspondence table between calibrated tests that would give the same shaped distribution for a given sample of examinees, but then it would no longer be true that an individual's score from either test was an unbiased estimate of the same "true score."

It is possible to estimate, from correctly -calibrated measures, characteristics of the "true-score" distribution such as its spread and proportions above cut-points. More complex statistical methods are required, however, because one must account for the spread of evidence about each of the observations, not just the best single-point estimates that are all that correspondence tables can convey. Under true-score test theory, classical formulas such as corrections for attenuation can be applied (e.g., Gulliksen, 1950/1987). Under item response theory, more specialized techniques such as those employed in NAEP are required (see Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992).

Calibration shares two important characteristics with equating. First, statistical and data-gathering procedures do not, in and of themselves, cause the relationship to come into being. Calibration is made possible only if tests gather information about the same skills and knowledge--generally, purposefully rather than by happenstance. Second, when the calibration requirements hold, the linking relationship does not depend in the main on other characteristics of the examinees. It is assumed (and must be verified in ways discussed below) that the concern is one of different distributions or amounts of "random noise" associated with observed scores, as measures of 'true scores." The kinds of verification noted above are carried out at the levels of both tests as a whole and individual items. For tests as a whole, one either checks whether essentially the same results are obtained when one fits calibration models separately for different groups (as defined for example by gender, ethnicity, age, or native language), or fits a single calibration model and looks for outliers associated with examinee characteristics. Checks at the item level are referred to in the IRT literature as Differential Item Functioning (DIF) studies (Holland & Wainer, 1993); they check for patterns of relationships among items that are, in the appropriate sense, consistent across different subgroups of examinees.

Calibration, Case 1: Different-length tests of the same skills

The simplest calibration situation in educational assessment differs from the equating situation in just one respect: Two tests are built to the same specifications, use similar items, and standard administration conditions, but use more or fewer items. The observed percent-correct scores from two tests built in this manner are, by construction, approximately calibrated in terms of expected percent correct on a hypothetical infinitely long test from the same specifications. (The item response theory methods discussed below could be used to fine-tune the calibration.) Unbiased estimates are obtained for individuals, but the short test would show a wider spread of scores than the long test. The short form would yield more errors about who was or was not above a given true-score level, but this is inherent in its sparser evidence and cannot be somehow calibrated or equated away. If equating formulas were inappropriately applied to make the long and short test have the same distributions of observed scores, then scores for the short test would tend to give people scores too close to the average, compared to their true scores.

Shooting free throws

Bob Linn (1993) uses the example of a basketball free-throw competition to illustrate calibration of varying-length tests. The coach wants to select players who shoot with at least 75% accuracy. The long test form is 20 tries; the short form is 4 tries. If a player's true accuracy is, say, 50%, her most likely outcome in either setting is 50%: 10 of 20, or 2 of 4. Accordingly, the properly calibrated estimate of the accuracy of a player who makes 50% of her shots is 50%. But because of the greater uncertainty associated with observing only 4 tries, a true 50% shooter has a probability of .31 of making at least 75% of her shots with the short form, but less than .01probability of making at least 75% of her shots with the long form.

An example concerning adult literacy tests

The shortened Survey Form of the TABE can be used to obtain estimates of performance in the three composite content areas of the full TABE, namely, reading, mathematics, and language, and of the total score (Pelavin Associates, 1994b, p. 30). It has the same proportional mix of items as the full TABE in all seven areas reported there (reading vocabulary, reading comprehension, mathematical computation, mathematical concepts/application, language mechanics, language expression, and spelling), but not enough in each to warrant separate subtest scores. The Survey Form would serve well for gauging an individual student's likely performance on the full battery, yielding an

estimate on the same scale but with less precision. But even though this would be true for every single student in a program, the proportions of students with scores beyond any designated point of interest on the scale (other than the group's mean happened to be) would be biased toward the extremes of the scale. For example, both the proportions of very high and very low performers would be overstated.

Calibration, Case 2: Item Response Theory (IRT)

Standard equating and the simple calibration situation described above depend on the construction of constrained observational settings. Inferences about behavior in two settings can be related because the settings are constructed to be similar. These approaches offer little guidance for tests intended to measure the same basic set of skills, but built to slightly different blueprints, such as harder or easier collections of similar items. Item response theory (IRT) lays out a framework that the interactions of students and items must exhibit to satisfy the axioms of "measurement" as it developed in the physical sciences (see, e.g., Rasch, 1960/1980). We confine the following discussion to right/wrong items, but extensions to ordered categories, counts, ratings, and other forms of data are available. *If* an IRT model is an adequate description of the patterns that occur in item responses, statistical machinery in the IRT framework enables one to "calibrate" tests in essentially the same sense as we calibrate physical instruments. In this case, the items and tests constructed from them are calibrated to the unobservable variable "tendency to get items of this kind right." The chances are best that this will happen for tests built from the same skills-and-format framework. The required regularities may also be found across sets of tasks written to different specifications, or from different conceptions of competence, but it is less likely. Whether it happens in a given application is an empirical question, to be answered with the data from an empirical study in which students are administered tasks from both assessments.

Stone-lifting as a measure of strength

Suppose we have a room full of stones, with weights from 5 to 500 pounds. To learn about people's strength in a way that corresponds to classical test theory, we would select a random sample of, say, 50 stones, and record how many a person can lift. This gives us (1) an estimate of the proportion of stones in the room an examinee could lift, if they tried them all, and (2) a way to rank people in terms of their strength—a norm-referenced inference. It doesn't tell us whether a particular person would be likely to lift a 100 pound stone—a criterion-referenced inference. A

different sample of 50 stones would give the same kind of information. Because one sample might contain more heavy stones, we could equate number-of-stones-lifted from the two samples of stones, just as we equate scores from alternate forms of the SAT or the TABE.

> Comparing peoples' strength in this manner requires strong people to lift very light stones, and the less strong to try heavy ones—both expending their time and energy without telling us much about their strength. A better system notes exactly which stones a person attempts. People can generally lift stones up through a certain weight, have mixed success in a narrow range, then rarely lift much heavier ones. We might characterize their strength by the point at which they have 50-50 chances of lifting a stone. Our accuracy for a particular person depends on how many stones are in the right range for him or her. Knowing Alice succeeds with 70 and 90, but not 110 and 130, narrows our estimate for her to "between 90 and 110." Adding 95, 100, and 105 pound stones makes her "test" more precise. The same stones tell us virtually nothing for measuring the strength of Jonathan, who has succeeded with 10 and 15 but not 20 or 25.
>
> If we haven't had a chance to weigh the stones ahead of time, we can observe, for a number of people, which stones they can lift and which they can't. Believing the stones are lined up in an order that applies in the same way to every person, we can use the relative frequencies with which stones are lifted to discover this order, then use it to measure more people for whom we believe the same ordering also makes sense.

Linking tests with IRT

In analogy to stone-lifting, IRT models for right/wrong test items propose that the probability that a person will respond correctly to an item is a function of (1) a parameter characterizing that particular person's proficiency and (2) one or more parameters characterizing that particular item's difficulty and perhaps other characteristics such as sensitivity. If the model holds, long and short tests, hard and easy ones, constructed from the pool of items might be "calibrated" like the stones.[2] The estimates for peoples'

[2] The likelihood functions for right and wrong responses to each of the items would be estimated from the data, characterizing the evidence each provided about proficiency on the domain of items. The likelihood

locations would be scattered around their "true" locations, with the degree of accuracy depending mainly on the number of items administered. Comparing peoples' measures with one another or with group distributions supports norm-referenced inferences about individuals. Comparing their scores with item locations supports criterion-referenced inferences. Peoples' IRT measures are approximately unbiased with long tests.[3]

Linking tests with IRT models amounts to estimating the locations of tasks from both tests, based on the responses of a sample of people who have taken at least some of each. *If* the IRT model fits the patterns in data well enough, then estimates of students' proficiency parameters are properly calibrated measures on the same scale. A correspondence table could be drawn up that matches the expected scores on the two tests for various values of "true" proficiency. The estimate of an individual student would have about the same expectation from either test, although measurement error might be small with one test and horrendously large with the other.

But the distribution of unbiased scores for individuals is generally not an unbiased estimate of the distribution of their true scores. The correspondence table described in the preceding paragraph gives the wrong answers about population characteristics such as dispersion, and the proportion of people above selected points on the scale. Statistical machinery is available (again, *if* the IRT model holds) to estimate population characteristics whether the test forms are long or short, hard or easy, but questions about group distributions must be addressed with data from the group as an ensemble. Mislevy, Beaton, Sheehan, and Kaplan (1992) describe how such procedures are used in NAEP. They show how using a demonstrably good estimate for every student in a group gives a demonstrably bad estimate about the characteristics of the group as a whole.

I have said that the benefits of calibration are achieved with IRT "*if* the model holds." Again, the *potential* characteristics of a linking between tests emanate from the kinds of skills the tests tap and the circumstances under which the observations are

---

function induced by a person's set of responses to several items would be the product of the appropriate ones of these. The highest or most probable point is his or her "maximum likelihood estimate" test score.

[3] They are "consistent" estimates, meaning that they approach being unbiased as the number of observations increases. In the interest of exposition, I will use the more familiar term "unbiased" loosely, to encompass "consistent."

made—not by statistical machinations after the observations have been made. The statistical machinations, if carried out properly, merely harness those relationships for use. Merely carrying out the machinations when the relationships have not been verified does not lead automatically to the properties associated with a kind of linkage. This is as true of IRT test calibration as any of linking approach.

We have learned over the past 20 years that IRT models tend to fit better if the items are more homogeneous, and the examinees are more homogeneous with respect to what the items require. For items, this includes not only their content, but the way in which they are presented—item formats, timing conditions, order of the items, and so on—and the uses to which they will be put, which affects student motivation. Estimates of group averages can change more from these causes than from a year's worth of schooling (Beaton & Zwick, 1990)! For students, differences in cultural backgrounds and educational experiences that interact with the item content can make it impossible to define a single common measure from a collection of items—the items tend to line up differently for members of the different groups. Merely standardizing the conditions of observation is not sufficient for "test scores" to support "measures." Exactly the same items, settings, and timings can lead to different characteristic patterns among students with different backgrounds, muddling the meaning of test scores and subverting comparisons among students.

Suppose, for example, the tasks in Test X are built around the skills Literacy Program X emphasizes, while the tasks in Test Y are built around the skills Literacy Program Y emphasizes. A combined test with both kinds of items might approximately "define a variable" among Program X students or Program Y students separately. The combined test used with all students together would produce *scores*, all right, and we could even construct parallel test forms and successfully equate them. But because of the *qualitative* difference between students' profiles of performance for X and Y items, the scores can *not* be thought of as measures on a single well-defined variable. We can attack this problem in two ways:

- Perhaps subsets of the tasks from the two tests assessments cohere, so that one or more portions of different assessments can be calibrated to variables that extend over groups.

• If an IRT model holds across all tasks within each student group separately, but not across groups, we might think of the tests as measures of two distinct variables rather than two alternative measures of the same variable. Rather than "calibrating" two tests to the same single variable, we would study the relationships among two (or more) distinct variables. This is the topic of the section on "projection."

Two examples concerning adult literacy tests

Pelavin Associates' (1994b) review of adult education tests states that "... the CASAS staff can create customized instruments, either by modifying an existing test to extend its range at the top or bottom, or by building a completely new test according to specification of achievement range, competencies, length, and desired accuracy. ... A total number correct for each subtest is converted, based on IRT difficulty calibrations and a Scale Score Conversion Chart, to CASAS scale scores" (p. 16). In doing so, the CASAS staff provides programs the option of more accurately estimating the proficiencies of their students by targeting the difficulty of customized forms. They also provide programs the option of reducing testing time by constructing short forms that give rougher approximations of the same CASAS scale scores. Programs exploiting these options sacrifice a degree of comparability to programs using other forms when it comes to estimating the distributions of their students on the CASAS scale, for purposes of evaluation and monitoring at the program level. Comparisons can still be made, though with the more complex methods described above rather than simply from summary statistics of students' observed scores.

The proficiency scales upon which TALS scores are reported were created in the 1985 Survey of Young Adult Literacy (SYAL; Kirsch & Jungeblut, 1986). An IRT model was designed to fit the 3105 SYAL person survey sample, and national distributions of scale scores were estimated. Because the purpose of SYAL was to estimate these population distributions, it was neither necessary nor cost-effective to present enough tasks to every respondent to obtain a precise estimates for each of them. Accordingly, a matrix-sample design was used, testing time was held down, and the more complex techniques mentioned above (e.g., Mislevy, 1991) were used to estimate population distributions. The TALS subtests consist of tasks of the same basic types, but enough are now presented to an examinee to obtain usable point estimates. These estimates can be interpreted in terms of the IRT scales established in the SYAL study through conversion tables provided by the publisher—but if accurate estimates of group distributions on these scales were desired, one

would resort again to the more complex estimation methods rather than using distributions of observed scores.

Calibration, Case 3: Judgments on an abstract proficiency scale

In recent years, IRT modeling has been extended from right/wrong items to ratings and partial-credit data (see, e.g., Thissen & Steinberg, 1986, for a taxonomy of models). It is sometimes possible to map judges' ratings of performance on complex tasks onto a common scale, from which properly calibrated measures of student performance might be derived. As with IRT for right/wrong items, an opportunity also exists to discover that performances over a given collection of tasks cannot be coherently summarized as values on a single variable! For reasons discussed below, linking in this manner is more likely to succeed for tasks sharing a focused conception of some aspect of competence.

The ACTFL language proficiency scales provide an example (recall the reading guidelines in Table 1). The process might be carried out within a single language or for performances across several languages, the latter representing a greater challenge. One or more judges, observing one or more performances of a student, would rate students' accomplishments in terms of ACTFL levels. Students would be characterized in terms of their tendencies to perform at levels of the guidelines, a variable at a higher level of abstraction than their actual sample of performances. Raters would be characterized in terms of their harshness or leniency. Interview topics or settings, possibly tailored and certainly interpreted individually for individual students, would be characterized in terms of their difficulty.

Even with successful training, some variation inevitably remains among judges as to the characteristics of any given performance. In a given setting, however, hard work and attention to unusual ratings can bring a judging system to what Deming calls "statistical control": the extent of variation among and within judges lies within steady, predictable ranges. The more narrowly defined a performance task is, the easier it will usually be for judges to come to agreement about the meanings of ratings, through words or examples, and the more closely their judgments will agree.

We can expect the typical amount of inter-judge variation in a system under control to increase as student competence is construed more abstractly, or as the range of ways it might be manifest broadens. This variation leads to larger measures of

uncertainty for students' scores (i.e., ratings induce more dispersed likelihood functions). A model such as Linacre's (1989) extension of IRT to individual raters and other facets of the judging situation helps identify "unusual" ratings—an interaction between a judge and a performance more out of synch with other ratings of that performance and other ratings by that judge than would be expected, given the usual distribution of variation among raters and performances. Relaying this information back to judges helps them come to agreement about criteria, and helps assure quality-control for high-stakes applications (for an example with Advance Placement Studio Art portfolio assessment, see Myford & Mislevy, 1995).

Experience suggests that, all other things being equal, the greater the degree of judgment demanded of raters, the more uncertainty associated with students' scores. The contrast with multiple-choice items can be dramatic. This latitude may be exactly what we want for instructing individual students, for it expands the richness, the variety, the individualization—the *usefulness*—of the assessment experience between the teacher and the student. But high-stakes applications in which a common framework of meaning is demanded over many students and across time may prompt us to develop more constrained guidelines; to break scoring into multiple, more narrowly-defined rating variables; to train raters more uniformly; or to increase the number of raters per performance, the number of performances per student, or both.

Comments on calibration

Educational assessments can be calibrated together if the evidence each conveys can be expressed in terms of a likelihood function on a common underlying variable. The expected score of a student would be the same in any of the assessments he or she can appropriately be administered, although there may be more or less evidence from different assessments, or for different students on the same assessment. One route to producing assessments that can be calibrated is to write them to blueprints that are the same except with more or fewer tasks. Alternatively, patterns of responses to a collection of multiple-choice or judgmentally-scored tasks may be satisfactorily approximated by an item response theory model. Arrangements of these tasks in different tests for different students or different purposes could then be calibrated in terms of a common underlying variable.

The proviso for the IRT route is that the patterns in data—interactions between students, tasks, and, if they are involved, judges—must accord with the regularities of the measurement model. Irregularities, such as when, compared to other tasks, some tasks are hard for some kinds of students but relatively easy for other kinds of students, are more likely to arise when we analyze more heterogeneous collections of tasks and students. IRT statistical machinery can be used to discover unwanted task-by-background interactions, explore the degree to which they impact inferences, and help determine whether breaking assessments into narrower sets of tasks accords better with the measurement paradigm. If the full range of assessment tasks don't, as a whole, conform to a measurement model, perhaps smaller, more homogeneous, groupings of tasks will.

Even when assessments do "measure the same thing" closely enough for our purposes, measuring it with different accuracy introduces complications for inferences at the level of groups of students. Although we get unbiased answers to questions about individual students and group means with "properly calibrated" scores, these same scores give biased estimates of population characteristics such as the spread and the proportion of students above a particular point on a scale. Statistical methods can provide the correct answers to these latter questions, but they are unfamiliar, complex, and address the configuration of the group's data as a whole rather than as a collection of individual scores for each student.

Projection

Basic concepts of projection

Suppose we have ascertained that two tests "measure different things." We can neither equate them nor calibrate them to a common frame of reference, but we may be able to gather data about the joint distribution of scores among relevant groups of students. The linking data might be either the performances of a common group of students that has been administered both assessments, or ratings of a common group of judges of performances from both assessments. We could then make *projections* about how a student with results from Assessment Y might perform on Assessment X, in terms

of probabilities of the various possibilities.[4] Projection uses data from a linking study to model the relationship among scores on the two assessments *and other characteristics of students that will be involved in inferences*. This phrase is highlighted because "measuring the same thing" in equating and calibrating situations meant we don't have worry about this—but now we do. The relationships among scores may change dramatically for, say, students in different instructional programs. As we will see, ignoring these differences can distort inferences and corrupt decisions.

If Ming Mei takes Test Y, what does this tell us about what her score might have been on Test X? While the statistical machinery required to carry it out and gauge its accuracy can be complex, the basic idea behind projection is straightforward: Administer both tests to a large sample of students "suitably similar" to Ming Mei. The distribution of X scores of the people with the same Y score as Ming Mei is a reasonable representation of how we think she might have done. The same idea applies when both X and Y yield multiple scores or ratings. The phrase "suitably similar" is the joker in this deck.

Two assessments can differ in many ways, including the content of the questions, the mode of testing, the conditions of administration, and factors affecting students' motivation. Because each of these changes impacts the constellations of skills, strategies, and attitudes each student brings to bear in the assessment context, some students will do better in one than another. In the projection approach to linking assessments, a sample of students is administered all of the assessments of interest (or interlocking portions of them) to estimate their joint distribution in that sample. From the resulting estimated distributions, we answer questions such as, "If I know Ming Mei's results on Assessment Y, what probability distribution represents my expectations for her on Assessment X?" The answer can be applied to an inference we'd have made using Assessment X results, had they been available. The spread of this projected distribution represents the added uncertainty that must be added to the usual measurement uncertainty associated with

---

[4] A best single guess is a "prediction." Predictions are sometimes used to link assessments, but if you carry out the analyses described here for projection, predictions fall out as a special result. I discuss the more enclusive approach because it is easier to understand in terms of the evidential value of observations for inferences.

performance within a given method—using just the best guess, the point prediction, would overstate our confidence. We can expect that the more differences there are between assessments' contents, methods, and contexts, the weaker the association between them will be.

The relationship between assessments can differ systematically for students with different backgrounds, different styles, or different ways of solving problems. Recall the math tests tailored to different math programs, the items of which could not be calibrated to a single underlying variable. The students instructed from one perspective will tend to do better with tasks written from the same perspective. When we speculate on how well Ming Mei would fare on Assessment X, we look at the X score distribution of "suitably similar" students with the same Assessment Y score(s) as hers. But suppose the distributions differ for students who have studied in Program X and those who have studied in Program Y. The answer to "What are my expectations for Ming Mei on Assessment X if I know her results on Assessment Y *and that she has studied in Program Y?*" will then differ considerably from the answer to "What are my expectations for Ming Mei on Assessment X if I know her results on Assessment Y *and that she has studied in Program X?*" An answer that doesn't take program study into account averages over the programs that the students in the linking sample happened to have taken, and gives a rather unsatisfactory statement about the competence of Ming Mei as an individual, for either a high-stakes decision about her accomplishments to date or a low-stakes decision to guide her subsequent instruction. We thus arrive at the first desideratum for a linking assessments through projection:

> 1. *A linking study intended to support inferences based on projection should include relationships among not only assessments, but other student variables that will be involved in the inference.*

A similar caveat in projection linking is that the relationships among assessments can change over time. This factor arises when test results are used for accountability purposes, to guide educational policy decisions. The actions that tend to maximally increase scores on one test may not be the same as those that increase scores on a different test, even though the relationship between students' scores on the two tests within any single time point may be highly related. Thus, high correlations among tests is not sufficient to treat them as interchangeable, if they tap different skills and these skills have different relationships to programs' instructional content.

For example, students who do well on multiple-choice questions about writing problems also tend to write well, compared with other students. At a given point in time, rankings from the two tests would rank the students similarly, and provide similar evidence for norm-referenced inferences. Now actually writing essays tends to raise the essay-writing skills of all students, even their ordering according to multiple-choice performances and according to essay performances remains similar. Suppose we estimate the joint relationship among essay and multiple-choice measures at a single point in time, and use only multiple-choice scores to track results over time. Our one-shot linking study is structurally unable to predict an interaction between assessments over time. No matter whether essay scores would have gone up or down relative to multiple-choice scores, it would project a given trend in multiple-choice scores to the same trend in essay scores. We thus arrive at the second desideratum for a linking assessments through projection:

> 2. *If projection is intended to support inferences about change and changes may differ for the different assessments involved, then linking studies need to be repeated over time to capture the changing nature of the relationships.*

Predicting marathon times

Bob Glover and Pete Schuder (1988) published a "predictor chart" for runners to approximate their times in races of different lengths, given results from any one of those lengths. There is a fairly consistent relationship between paces per mile for all distances, they note. "The two basic formulas are: to predict your 10K time from your 5K time, multiply by 2 and add 1 minute; to predict your marathon time [26 miles] from your half marathon time, multiply by 2 and add 10 minutes" (p. 589). The predictions from correspondence tables built from these relationships are known to be better estimates when the length predicted and the length known are similar; since similar factors are at work for similar length races, the relationships among similar distances are essentially calibrations. For long distances, however, not only do the predictions become less accurate, but the errors of predictions are found to be systematically related to other identifiable factors. Glover and Schuder note, for example, that "the predicted marathon times for women may be inaccurate (slow) by up to five minutes. Generally women can run 1-2 minutes faster than listed on the chart for the marathon" (1988, p. 589). This dependence of the relationship between measures upon such factors indicates that we are dealing with projection rather than calibration, and that predictions which ignore the associated variables will lead to systematic errors.

Some connections with adult literacy tests

Test documentation indicates that the TABE has been administered to a large sample of GED test takers in order to permit prediction of GED success from TABE performance. The predictions are "best estimates" from predictive distributions. The TABE documentation cautions that these relationships were established on a sample of candidates who had prepared to take the GED, and might not hold for the general population. This caveat properly recognizes that among a group of students obtaining the same TABE score, those who have prepared for the specific materials and question types the GED requires are more likely to do well on the GED than those who haven't.

Predictions of this type are fairly familiar, and are widely used to help individuals with diagnosis and planning. As noted above, however, the distribution of these point predictions for a *group* of students will understate the spread of the group's scores in the predicted test—even if the predictions are correct on the average. To properly approximate the projected score distribution on the target test requires using the entire predictive functions of individuals, not just the single best point estimates. Recently, Bloxom, Pashley, Nicewander, and Yan (1995) illustrated how this can be done by projecting members of the armed services' distributions on the National Assessment of Educational Progress, given their performance on the Armed Services Vocational Aptitude Battery (ASVAB) and a variety of background variables.

Comments on projection

Projection can be carried out with assessments constructed around different conceptions of students' competence, or around the same conceptions but with tasks that differ in format or content. Such assessments provide qualitatively different evidence for various conjectures about the competence of groups or individuals. With care, it is possible to estimate the joint relationships among the assessment scores and other variables of interest in a linking study. The results can be used make projections about performance on one assessment from observed performance on the other, in terms of a distribution of possible outcomes. Although it is possible to get a point prediction, no simple one-to-one correspondence table captures the full import of the link for two reasons: (1) using only the best estimate neglects the uncertainty associated with the projection, and therefore of inferences about individual students, and (2) the relationships,

and therefore inferences they inform, can vary substantially with students' education and background, and can change with the passage of time and instructional interventions.

While the statistical machinery exists to characterize the relationship among different tests in specified contexts and make defensible projections, the enterprise demands circumspection in the following ways:

1. When the contexts change, the relationships among test scores can change too—sometimes in ways that undermine the very inferences that are most important. Evaluating program effects with tests that are differentially related to the instructional content of the program is a prime example. The more assessments arouse different aspects of students' knowledge, skills, and attitudes, the wider the door opens for students to perform differently in the different settings.

2. Moderate associations among assessments can support inferences about how *groups* of students might fare under different alternatives. Thus projection can be useful even when correlations between tests are relatively low. But projections for *individual* students in high-stakes applications would demand not only strong empirical relationships, but vigorous efforts to identify groups (through background investigations) and individuals (through additional sources of information) for whom the usual relationships fail to hold.

Statistical Moderation

"Moderation" is relatively new term in educational testing, although the problem it addresses is not. The goal is to match up scores from different tests which admittedly do not "measure the same thing." It differs from projection in the following sense: While projection attempts to characterize the evidence a performance on one assessment conveys about likely performance on another, moderation simply asks for a one-to-one matchup among assessments as to their worth. Moderation is thus an evaluation of value, as opposed to an examination of evidence. We consider two classes of linking procedures that have evolved to meet this desire. The methods of *statistical moderation* have been developing for more than fifty years as a subclass of scaling procedures (see, for example, Angoff, 1984, and Keeves, 1988). They are statistical in appearance, using estimated score distributions to determine score levels that are deemed comparable.

Social moderation, a more recent development, relies upon direct judgments. It will be discussed in a following section.

Basic concepts of statistical moderation

Statistical moderation aligns score distributions in essentially the same manner as equating, but now with tests that admittedly do not "measure the same thing." If two tests can sensibly be administered to students, statistical moderation simply applies the *formulas* of equating, without claiming a measurement *justification*. If two tests have in fact been constructed to the same blueprint, applying equating formulas really does equate them. If they haven't been constructed to the same blueprint, applying the same formulas renders them statistically moderated. Let us explore what this means.

The scales of the SAT Verbal and Quantitative scores (SAT-V and SAT-Q) illustrate this simple case. In April 1941, 10,654 students took the SAT. Their SAT-V and SAT-Q formula scores (number correct, minus a fraction of the number wrong) were both transformed to an average of 500 and a standard deviation of 100. Tables of correspondence thus mapped both SAT-Q and SAT-V formula scores into the same 200-to-800 range. An SAT-V scale score and an SAT-Q scale score with the same numerical value are "comparable" in this normed-referenced sense: In the 1941 sample, the proportion of examinees with SAT-V scores at or above this level, and the proportion with SAT-M scores at or above this level, were the same. (Whether similar proportions of this year's sample have these scores is quite a different question, and in general they don't.).

A more complex version of statistical moderation introduces "moderator tests." Moderator tests are a device for linking disparate "special" assessments that are taken by students in different programs or jurisdictions, or for different reasons—for example, German tests for students who study German, and Physics tests for students who study Physics. Two scores are obtained from each student in a linking study: one for the appropriate special assessment, and one on a "moderator test" that all students take. Scores on the moderator test are used to match up performance on the special tests. The rationale is articulated in *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests* (Donlon, 1984, p. 21). An analogy of how this would look in an athletic context follows.

Using Mile-Run Times to Moderate Batting Averages and Goals-Scored

To "moderate" baseball batting averages and hockey goals-scored using mile-run times, we would first obtain batting averages and run times from a group of baseball players, and goals-scored and run times from a group of hockey players. Within the baseball players, the best batting average is matched to the best run times, the average to the average, and so on; similar for goals-scored and run times within hockey players. To find out what number of goals-scored is "comparable" to a batting average of .250, we (1) find out what proportion of the baseball players in the baseline group hit below this average; say it's 60%; (2) look up the mile-run time that 60% of the baseball players were slower than; say its 5:10; (3) look up how many of the hockey players were slower than 5:10 in the mile; say its 30%; (4) look up how many goals-scored that 30% of the hockey players were below; say its 22.

Suppose hockey players tend to be faster runners than baseball players. Figure 1 shows the linkages that result from goals-scored to running times, and running times to batting averages. This procedure would be consistent with an argument that running times measure athletic ability, and hockey players, because they run so fast, would have high batting averages if they were baseball players. The judgment of the relative value of batting and goal-scoring skills thus lies implicit in the choice of the moderator test.

Notice that the three disparate tests of aspects of athletic ability (mile-run times, batting averages, and goals-scored) will be differentially affected by clinics in running, batting, and hockey skills. Considerable improvement in hitting would be reflected only in batting averages, so using mile-run times as a measure of program effect would miss important outcomes of the batting clinic. If only batting averages were measured, the improvement could be misconstrued as predicting commensurate gains in run-times and hockey skills.

[[Figure 1: mile-run moderation]]

Statistical moderation and the estimation of program effects

When two statistically-moderated tests were constructed to tap different mixes of skills, different samples of people can give substantially different linking functions. A given X score and Y score might be matched with respect to a set of suburban high school students, but the X score might come out much higher for urban high school students while the Y score came out higher for adult ESL students. This too is a matter for empirical verification. The more similar the linking function comes out for different

relevant subsamples, the more justification one has for treating such a link as a calibration: both tests can be construed as providing information about the same underlying mix of skills, though perhaps with differing amounts and distributions of measurement error.

An especially important check in this respect is pre- and post-test samples, for statistical moderation linking is especially sensitive to the effects of instruction. Any two tests (TABE scores and shoe sizes!) can be statistically moderated, in the sense that their respective distributions can be aligned with respect to some group of people. There is no guarantee, however, that skills improvements that students make in educational programs will be equally reflected in the "moderated" scores. The more discrepant the tests are with respect to the skills they tap, the more disagreements as to measured gains can be—and the greater they will be as the instruction in a program is more closely related to the skills set of one test than the other.

A subtle variation of statistical moderation is applying an IRT model across collections of tasks and students in the face of the unwanted interactions we discussed in the Calibration section. If a composite test consisting of tasks keyed to Program X and Program Y are calibrated together with responses from students from both programs, task-type may interact with student-program. Consequential interactions drive us from calibration into the realm of statistical moderation. Comparisons among students and between groups could turn out differently with a different balance of task types or a different balance of students in the calibration group. If the interaction is inconsequential, however, inferences may proceed under the conceit of a common measurement model; few would differ if the balance of items or the composition of the calibration sample were to shift. Determining the sensitivity of a link to these factors is a hallmark of a responsible IRT linking study.

Some connections with adult literacy tests

A first, familiar, example of statistically moderated tests is afforded by "grade equivalence" scales of standardized reading achievement tests. Any test of some mix of reading skills can be administered to a sample of students from each grade in school, and a grade-equivalents scale can be established for its raw scores. If two reading tests tap nearly the same mix of skills, the resulting grade equivalent scales may approach the properties of "calibrated" scales (see the section above on calibration), but again this is a

function not of the statistical machinations but of the structuring of the tasks, which for certain tests might happen to match up well. Only in these cases will average gains as measured by either test tend to be similar for groups.

A second, prospective, example, is the notion of linking diverse tests to a single standard test. This idea has been suggested in connection with the notion of "measuring national standards with locally-developed tests," by linking all of them to, say, the National Assessment for Educational Progress. This is statistical moderation with respect to a common moderator test. The equivalent in adult education literacy tests might be linking through, say, the NAEP reading or SYAL literacy scales, which might be considered as a standard by virtue of nationally-representative surveys in their terms.

Were this done, variations as to linking samples and moderating tests would have no effect on norm-referenced comparisons of performances *within* specific tests, but they could affect markedly norm-referenced comparisons from one specific test to another. Carrying out the procedures of statistical moderation with different samples of students, at different points in time, or with different moderator tests can produce markedly different numerical links among tests. Particular choices of these variables can be specified as an operational definition of comparability, but moderated scores in and of themselves offer no clue as to how much the results would differ if the choices were altered. The more the tests vary as to content, format, or context, the more the results would vary under alternative moderation schemes. A sensitivity study compares results obtained under different alternatives, revealing which inferences are sensitive to the choices. We would have little confidence in a comparison of, say, subgroup means across "moderated" test scores unless it held up under a broad range of choices for linking samples and moderator tests.

In an application that uses a moderating test, its specification determines the locus of value for "comparable worth." In the preceding frivolous sports example, baseball players' performances were assigned lower values than hockey players' goals-scored, simply because hockey players ran faster. In a serious educational context, consider using statistical moderation to link disparate educational assessments from clusters of schools through NAEP/SYAL. To the degree that focusing on skills not emphasized in NAEP/SYAL trades off against skills that are, this arrangement would work in the favor of clusters whose tests were most closely aligned with NAEP/SYAL, and against clusters as their content and methodology departed from it.

Social Moderation

Basic concepts of social moderation

Social moderation calls for direct judgments about the comparability of levels of performances on different assessments. This process could apply to performances in assessment contexts that already require judgment, or score levels in objectively-scored tests. As an example, Wilson (1992) describes the "verification" process in Victoria, Australia. Samples of students' performances on different assessments in different localities are brought together at a single site to adjudicate comparable levels of performance.

Auxiliary information on common assessment tasks may be solicited to supplement direct judgment. For example, the assessments of two school districts might contain tasks unique to each, but also a common core present in both. It would then be possible to compare the score distributions on unique tasks of students from both districts who had the same levels of performance on the common tasks. If nothing more were done but to simply align these distributions, we would have an instance of statistical moderation. If judgments were used to adjust the resulting matchup on the basis of factors left out of statistical moderation, such as the relevance of the common tasks to the unique tasks, we would have social moderation.

Like statistical moderation, social moderation is founded upon a particular definition of comparability, defined in terms of a given process, at a given point in time, with a given set of people (now the social moderators, corresponding to the linking sample of students). No built-in mechanism indicates the uncertainties associated with these choices. Like statistical moderation, social moderation can also be supported by sensitivity studies. How much do the alignments change if we carry them out in different parts of the country? With teachers versus content-area experts, or students, or community members carrying out the judgments? With different supplemental information to aide the processes? Again, we would have little confidence in inferences based on moderated test scores unless they held up under a broad range of reasonable alternatives for carrying out the moderation process.

Scoring the Decathlon

The decathlon is a medley of ten track and field events: 100-meter dash, long jump, shot put, high jump, 400-meter run, 110-meter hurdles, discus throw, pole vault, javelin throw, and 1500 meter run. Conditions are standardized within events, and it is easy to rank competitors' performances within each. To obtain an overall score, however, requires a common scale of value. This is accomplished by mapping each event's performance (a height, a time, or a distance) into a 0-1000 point scale, where sums are accumulated and overall performance is determined. A table was established in 1912 for the decathlon's first appearance in the Olympics by the International Amateur Athletic Federation (IAAF) by a consensus among experts. Performances corresponding to then-current world records were aligned across events, and lesser performances were awarded lower scores in a manner the committee members judged to be comparable.

The IAAF has revised the decathlon scoring table in 1936 and 1950 to reflect improvements in world-class performance and to reflect different philosophies of valuation. All of these earlier tables emphasized excellent performance in individual events, so that a superior performance in one event could more than offset relatively poor showings in several others. By scaling down the increases at the highest levels of performance, revisions in 1964 and 1985 favored the athlete who could perform well in many events.

Table 4 shows the performance and total scores of the top eight contenders from the 1932 Olympics in Los Angeles. Their scores under both the 1912 tables, which were in effect at the time, and the 1985 tables are included. We notice first that the 1985 totals are all lower, reflecting the fact that the top performances in 1932 were not as impressive when judged in the competitive milieu of 1985. Moreover, James Bausch's performances won the gold medal, but he would have finished behind Akilles Järvinen had the 1985 tables been used. Bausch's outstanding shot put distance more than compensated for his relatively slower running times.

[[Table 4: 1932 Olympics Decathlon results]]

Some connections with adult literacy tests

Two applications of social moderation can be mentioned in the arena of adult basic education, one extant and the other potential. The existing example is the GED. Passing the GED test to receive the GED certificate is claimed to be equivalent in some sense to graduating high school. And it is probably true that candidates who pass the GED have many reading, computation, and knowledge capabilities on a par with students

who would graduate from typical high schools, even though the actual skills and accomplishments required for the two routes differ in important essentials.

The prospective example of social moderation would be to mediate observed performance in different tests through the ACTFL reading guidelines (Table 1). It might be possible for observers to come to a socially determined agreement as to performances on different tests that represent, say, novice or advanced English reading skills. Without further data, however, little claim might be warranted for how a person judged to be advanced with respect to one test might fare on another. Confusion can result if different programs elected to evaluate their students only with respect to tests well matched to their instruction, because students would be classified higher on the tests associated with their particular program than with any other tests. A student might thus be "advanced" only with respect to one special testing regime, and not on any other. A "Lake Wobegone" effect would thus appear.

Some comments on moderation

Moderation should not be viewed as an application of the principles of statistical inference, but as a way to specify "the rules of the game." It can yield an agreed-upon way of comparing students who differ qualitatively, but it doesn't make information from tests that aren't built to measure the same thing, function as if they did. An arbitrarily determined operational definition of comparability must be defended. In *statistical* moderation, this means defending the reasonableness of the linking sample and, if there is one, the moderating test. It is bolstered by a sensitivity study showing how inferences differ if these specifications change to other reasonable choices. In *social* moderation, consensual processes for selecting judges and carrying out the alignment constitutes a first, necessary, line of defense, which can similarly be bolstered by sensitivity studies.

Conclusion

An educational assessment consists of opportunities for students to use their skills and apply their knowledge in content domains. A particular collection of tasks evokes in each student a unique mix of knowledge, skills, and strategies. Results summarize their performances in terms of a simplified reporting scheme, in a framework determined by a conception of important features of students' competence. If we construct assessments carefully and interpret them in light of other evidence, they can provide useful evidence about aspects of students' competencies to guide decisions about instruction and policy.

In any skill area defined as broadly as "literacy skills," developing competence has many aspects. No single score can give full picture of the range of competencies of very different kinds of students in different instructional programs. Accordingly, multiple sources of evidence—different types of formats, contexts, and skill demands—must be considered. Some of these will be broadly meaningful and useful; others will be more idiosyncratic in their utility, at the level of the state, the program, or the individual student.

No simple statistical machinery can somehow transform the results from any two arbitrarily selected assessments so that they provide interchangeable information about any question one might raise. Such strong linkage is in fact accomplished routinely when alternative forms of the SAT or of the ASVAB are equated—but only because these forms are written to the same tight form and content constraints, and administered under standard conditions. The simple and powerful linking achieved in these special cases emanates not from the statistical procures used to map the correspondence, but from the way the assessments are constructed.

What, then, can we say about prospects for linking disparate assessments in a national system of assessments? First, it *isn't* possible to construct once-and-for-all correspondence tables to "calibrate" whatever assessments might be built in different programs or states to provide different kinds of information about students. What *is* possible, with carefully-planned continual linking studies and less familiar statistical methodology, is attaining the less ambitious, but more realistic, goals:

- *Comparing directly levels of performance across programs in terms of common "universal" indicators of performance on a market basket of consensually-defined tasks in standard conditions.* Some aspects of competence and assessment contexts for gathering evidence about them will be considered useful by a wide range of programs, and components of an assessment system can solicit information in about them in much the same way for all (see sections on Equating and Calibration, Case 2). A role like this might be envisaged for something like selected scales from the National Assessment or the Survey of Young Adult Literacy. Gathering of information in this standardized manner, however, can fail to provide sufficient information about aspects of competence holding different salience for different programs. Common components should thus be seen as but one of many sources of evidence about students' competencies—and a limited one at that, as far as providing evidence about the variety of individual competencies that

characterize a nation of students. These kinds of assessments—and in particular pre-post comparisons with these assessments—would provide seriously incomplete information to evaluate the effectiveness of programs, to the extent that their focus did not match program objectives.

- *Estimating levels of performance of groups or individuals within clusters of programs with similar objectives—possibly in quite different ways in different clusters—at the levels of accuracy demanded by purposes within clusters, with common assessments focused on those objectives.* These components of programs' assessments might gather evidence for different purposes, types of students, or levels of proficiency, to complement information gathered by "universal" components.

- *Making projections about how students from one program might have performed on the assessment of another.* When students can be administered portions of different clusters' assessments under conditions similar to those in which they are used operationally, we can estimate the joint distribution of results on those assessments. We can then address "what if" questions about what performances of groups or individuals who took one of the assessments might have been on the other (see the section on Projection). The more the assessments differ as to their form, content, and context, though, the more uncertainty is associated with these projections; the more they can be expected to vary with students' background and educational characteristics; the more they can shift over time. Unless very strong relationships are observed, and found to hold up over time and across different types of students, high-stakes uses for individuals through this route are hazardous.

References

American Council on the Training of Foreign Languages. (1989). *ACTFL proficiency guidelines.* Yonkers, NY: Author.

Angoff, W.H. (1984). *Scales, norms, and equivalent scores.* Princeton: Educational Testing Service.Beaton & Zwick, 1990)

Bloxom, B., Pashley, P.J., Nicewander, W.A., & Yan, D. (1995). Linking to a large-scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics*, *20*, 1-26.

Cohen, S.H., & Cohen, M.J. (1985). Slosson Oral Reading Test. In D.J. Keyser & R.C. Sweetland (eds.), *Test Critiques,* Vol. IV. Austin, TX: Pro-Ed.

Davidowitz, S. (1977). *Betting thoroughbreds: A professional's guide for the horseplayer.* New York: E.P. Dutton.

Deming. W.E. (1980). *Scientific methods in administration and management.* Course No. 617. Washington, DC: George Washington University.

Development Associates. (1992). First interim report: National evaluation of adult education programs. Arlington, VA: Author.

Donlon, T.F. (Ed.) (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board.

Edgeworth, F.Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51,* 599-635.

Edgeworth, F.Y. (1892). Correlated averages. *Philosophical Magazine,* 5th Series, *34*, 190-204.

Glover, B., & Schuder, P. (1988). *The new competitive runner's handbook.* New York: Penguin Books.

Green, B. (1978). In defense of measurement. *American Psychologist, 33*, 664-670.

Gulliksen, H. (1950/1987). *Theory of mental tests.* New York: Wiley. Reprint, Hillsdale, NJ: Erlbaum.

Holland, P.W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Jastak, S., & Wilkinson, G.S. (1984). *WRAT-R Administration Manual.* Wilmington, DE: Jastak Associates.

Keeves, J. (1988). Scaling achievement test scores. In T. Husen & T.N. Postlethwaite (Eds.), *International Encyclopedia of Education.* Oxford: Pergamon Press.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* Princeton, NJ: National Assessment of Educational Progress/Educational Testing Service.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement.* Chicago: MESA Press.

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

Lord, F.M. (1962). Estimating norms by item sampling. *Educational and Psychological Measurement*, *22*, 259-267.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Mislevy, R.J. (1993). *Linking educational assessments: Concepts, issues, methods, and prospects.* (foreword by R.L. Linn) Princeton, NJ: Policy Information Center, Educational Testing Service. (ERIC #: ED-353-302)

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133-161.

Myford, C.M., & Mislevy, R.J. (1995). *Monitoring and improving a portfolio assessment system.* Center for Performance Assessment Research Report. Princeton, NJ: Educational Testing Service.

Pelavin Associates. (March 25, 1994a). *Comparing adult education tests: A meeting of experts.* Washington, D.C.: Author.

Pelavin Associates. (March 25, 1994b). *Comparing adult education tests: Test descriptions.* Washington, D.C.: Author.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 221-262). New York: American Council on Education/Macmillan.

Porter, A. C. (1991). Assessing national goals: some measurement dilemmas. In T. Wardell (Ed.),*The assessment of national goals. Proceedings of the 1990 ETS Invitational Conference* (pp. 21-42). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research/Chicago: University of Chicago Press (reprint).

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 263-331). New York: American Council on Education/Macmillan.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-292.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-577.

Wallechinsky, D. (1988). *The complete book of the Olympics*. New York: Viking Penguin.

Wilson, M.R. (1992). *The integration of school-based assessments into a state-wide assessment sytem: Historical perspectives and contemporary issues*. Unpublished manuscript prepared for the California Assessment Program. Berkeley, CA: University of California, Berkeley.

List of Tables

List of Figures

Table 1

ACTFL Proficiency Guidelines for Reading*

Level	Generic Description
Novice-Low	Able occasionally to identify isolated words and/or major phrases when strongly supported by context.
Novice-Mid	Able to recognize the symbols of an alphabetic and/or syllabic writing system and/or a limited number of characters in a system that uses characters. The reader can identify an increasing number of highly contextualized words and/or phrases including cognates and borrowed words, where appropriate. Material understood rarely exceeds a single phrase at a time, and rereading may be required.
Novice-High	Has sufficient control of the writing system to interpret written language in areas of practical need. Where vocabulary has been learned, can read for instructional and directional purposes standardized messages, phrases, or expressions, such as some items on menus, schedules, timetables, maps, and signs. At times, but not on a consistent basis, the novice-high reader may be able to derive meaning from material at a slightly higher level where context and/or extralinguistic background knowledge are supportive.
Intermediate-Low	Able to understand main ideas and/or some facts from the simplest connected texts dealing with basic personal and social needs. Such texts are linguistically noncomplex and have a clear underlying internal structure, for example chronological sequencing. They impart basic information about which the reader has to make only minimal suppositions or to which the reader brings personal interest and/or knowledge. Examples include messages with social purposes or information for the widest possible audience, such as public announcements and short, straightforward instructions for dealing with public life. Some misunderstandings will occur.
Intermediate-Mid	Able to read consistently with increased understanding simple connected texts dealing with a variety of basic and social needs. Such texts are still linguistically noncomplex and have a clear underlying internal structure. They impart basic information about which the reader has to make minimal suppositions and to which the reader brings personal information and/or knowledge. Examples may include short, straightforward descriptions of persons, places, and things, written for a wide audience.

Table 1, continued

ACTFL Proficiency Guidelines for Reading

Level	Generic Description
Intermediate-High	Able to read consistently with full understanding simple connected texts dealing with basic personal and social needs about which the reader has personal interest and/or knowledge. Can get some main ideas and details from texts at the next higher level featuring description and narration. Structural complexity may interfere with comprehension; for example, basic grammatical relations may be misinterpreted and temporal references may rely primarily on lexical items. Has some difficulty with cohesive factors in discourse, such as matching pronouns with referents. While texts do not differ significantly from those at the Advanced level, comprehension is less consistent. May have to read several times for understanding.
Advanced	Able to read somewhat longer prose of several paragraphs in length, particularly if presented with a clear underlying structure. The prose is predominantly in familiar sentence patterns. Reader gets the main ideas and facts and misses some details. Comprehension derives not only from situational and subject matter knowledge but from increasing control of the language. Texts at this level include descriptions and narrations such as simple short stories, news items, bibliographical information, social notices, personal correspondence, routinized business letters, and simple technical material written for the general reader.
Advanced-Plus	Able to follow essential points of written discourse at the Superior level in areas of special interest or knowledge. Able to understand parts of texts which are conceptually abstract and linguistically complex, and/or texts which treat unfamiliar topics and situations, as well as some texts which involve aspects of target-language culture. Able to comprehend the facts to make appropriate inferences. an emerging awareness of the aesthetic properties of language and of its literary styles permits comprehension of a wider variety of texts, including literary. Misunderstandings may occur.

Table 1, continued

ACTFL Proficiency Guidelines for Reading

Level	Generic Description
Superior	Able to read with almost complete comprehension and at normal speed expository prose on unfamiliar subjects and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on the knowledge of the target culture. Reads easily for pleasure. Superior-level texts feature hypotheses, argumentation, and supported opinions, and include grammatical patterns and vocabulary ordinarily encountered in academic/professional reading. At this level, due to the control of general vocabulary and structure, the reader is almost always able to match the meanings derived from extralinguistic knowledge with meanings derived from knowledge of the language, allowing for smooth and efficient reading of diverse texts. Occasional misunderstandings may still occur; for example, the reader may experience some difficulty with unusually complex structures and low-frequency idioms. At the superior level the reader can match strategies, top-down or bottom-up, which are most appropriate to the text. (Top-down strategies rely on real world knowledge and prediction based on genre and organizational scheme of the text. Bottom-up strategies rely on actual linguistic knowledge.) Material at this level will include a variety of literary texts, editorials, correspondence, general reports, and technical material in professional fields. Rereading is rarely necessary, and misreading is rare.
Distinguished	Able to read fluently and accurately most styles and forms of the language pertinent to academic and professional needs. Able to relate inferences in the text to real-world knowledge and understand almost all sociolinguistic and cultural references by processing language from within the cultural framework. Able to understand the writer's use of nuance and subtlety. Can readily follow unpredictable turns of thought and author intent in such materials as sophisticated editorials, specialized journal articles, and literary texts such as novels, plays, poems, as well as in any subject matter area directed to the general reader.

* Based on the *ACTFL proficiency guidelines,* American Council on the Training of Foreign Languages (1989).

Table 2

Methods of Linking Educational Assessments

Link	Description	Procedure	Example	Comments
Equating	Equated scores from tests taken to provide equivalent evidence for all conjectures. Score levels *and weights of evidence* match up between scores on tests.	1. Construct tests from same blueprint. 2. Estimate distribution of tests in given population. 3. Make correspondence table that makes distributions match.	Two forms of a driver's license test, written to the same content and format specifications.	Foundation is not statistical procedure, but the way tests are constructed.
Calibration	Tests "measure the same thing," but perhaps with different accuracy or in different ways. Results from each test are mapped to a common variable, matching up the most likely score of a given student on all tests.	*Case 1:* Use same content, format, & difficulty blueprint to construct tests, but with more or less items on different tests. Expected percents correct are calibrated. *Case 2:* Construct tests from a collection of items that fits an IRT model satisfactorily. Carry out inferences in terms of IRT proficiency variable. *Case 3:* Obtain judgements of performances on a common more abstractly defined variable. Verify consistency of judgements (varieties of statistical moderation).	*Case 1:* A long form and a short form of an interest inventory questionnaire. *Case 2:* NAEP Geometry subscale for grades 4 and 8, connected by IRT scale with common items. *Case 3:* Judges' ratings of AP Studio Art portfolios, with student-selected art projects.	Correspondence table matches up "best estimates," but because weights of evidence may differ, the distribution of "best estimates" can differ over tests. Same expected point estimates for individual students, but with differing accuracy. Different estimates of many group characteristics, e.g., variance & population proportion above cut point.

(continued)

Table 2 (continued)

Link	Description	Procedure	Example	Comments
Projection	Tests don't "measure the same thing," but can estimate the empirical relationships among their scores. Observing score on Y, can calculate what you'd be likely to observe if X were administered.	Administer tests to the same students, and estimate joint distribution. Can derive predictive distribution for Test X performance, given Test Y observation. Can do conditional on additional information about student.	Determine joint distribution among students' multiple choice science scores, lab notebook ratings, and judgements of observed experimental procedures.	What Test Y tells you about what Test X performance might have been can change with additional information about a student. Estimated relationships can vary with the group of students in the linking study and over time in ways that distort trends and group comparisons.
Statistical moderation	Tests don't "measure the same thing," but can match up distributions of their scores in real or hypothetical groups of students to obtain correspondence table of "comparable" scores.	*Case 1:* If can administer both X and Y to same students, *estimate* X & Y distributions. Align X and Y with equating formulas. *Case 2:* If not, administer X and "moderator" Assessment Z to one group, and Z and Y to another. *Impute* X & Y distributions for hypothetical common group. Use formulas of equating to align X and Y.	*Case 1:* Correspondence table between SAT and ACT college entrance exams, based on students who took both. *Case 2:* Achievement results from History, Spanish, and Chemistry put on "comparable" scales, using common SAT-V and SAT-M tests as moderators.	Comments for projection also apply to statistical moderation. "Comparable" scores need not offer comparable evidence about nature of students' competence. Rather, they are perceived to be of comparable value in a given context, for a given purpose.
Social moderation	Tests don't "measure the same thing," but can match up distributions by direct judgment to obtain correspondence table of "comparable" scores.	Obtain samples of performances from two assessments. Have judges determine which levels of performance on the two are to be treated as comparable. Can be aided by performance on a common assessment.	Obtain samples of Oregon and Arizona essays, each rated through their own rubrics. Determine, through comparisons of scores given to examples, "comparable" levels of score scales.	"Comparable" scores need not offer comparable evidence about nature of students' competence. They are perceived to be of comparable value in a given context, for a given purpose.

Table 3

Parallel Time Chart for $1\frac{1}{8}$ Mile Races at Belmont and Aqueduct*

Speed Figure	Belmont (one turn)	Aqueduct (two turns)
136	1:46 2/5	1:48 1/5
132	1:46 4/5	1:48 3/5
128	1:47 1/5	1:49
124	1:47 3/5	1:49 2/5
120	1:48	1:49 4/5
116	1:48 2/5	1:50 1/5
112	1:49	1:50 4/5
108	1:49 2/5	1:51 1/5
104	1:50	1:51 4/5
100	1:50 3/5	1:52 1/5
96	1:50 4/5	1:52 3/5
92	1:51 1/5	1:53
88	1:51 3/5	1:53 2/5
84	1:52 1/5	1:54
80	1:52 3/5	1:54 2/5
76	1:53 1/5	1:55
72	1:53 3/5	1:55 2/5
68	1:54	1:55 4/5
64	1:54 2/5	1:56 1/5
60	1:55	1:56 3/5

* Based on Davidowitz, 1977, p. 204.

Table 4

Decathlon Results from the 1932 Olympics*

Name		Throwing Events			Jumping Events				Running Events			Total Points	
		Disc	SP	Jav	HJ	LJ	PV	110H	100 M	400 M	1500M	1912 Table	1985 Table
1. James Bausch	USA	44.58	15.32	61.91	1.70	6.95	4.00	16.2	11.7	54.2	5:17.0	8462	6735
2. Akilles Järvinen	FIN	36.80	13.11	61.00	1.75	7.00	3.60	15.7	11.1	50.6	4:47.0	8292	6879
3. Wolrad Eberle	GER	41.34	13.22	57.49	1.65	6.77	3.50	16.7	11.4	50.8	4:34.4	8031	6661
4. Wilson Charles	USA	38.71	12.56	47.72	1.85	7.24	3.40	16.2	11.2	51.2	4:39.8	7985	6716
5. Hans-Heinrich Sievert	GER	44.54	14.50	53.91	1.78	6.97	3.20	16.1	11.4	53.6	5:18.0	7941	6515
6. Paavo Yrölä	FIN	40.77	13.68	56.12	1.75	6.59	3.10	17.0	11.8	52.6	4:37.4	7688	6385
7. Clyde Clifford Coffman	USA	34.40	11.86	48.88	1.70	6.77	4.00	17.8	11.3	51.8	4:48.0	7534	6265
8. Robert Tisdall	IRL	33.31	12.58	45.26	1.65	6.60	3.20	15.5	11.3	49.0	4:34.4	7327	6398

- From Wallechinsky, 1988, *The complete book of the Olympics*

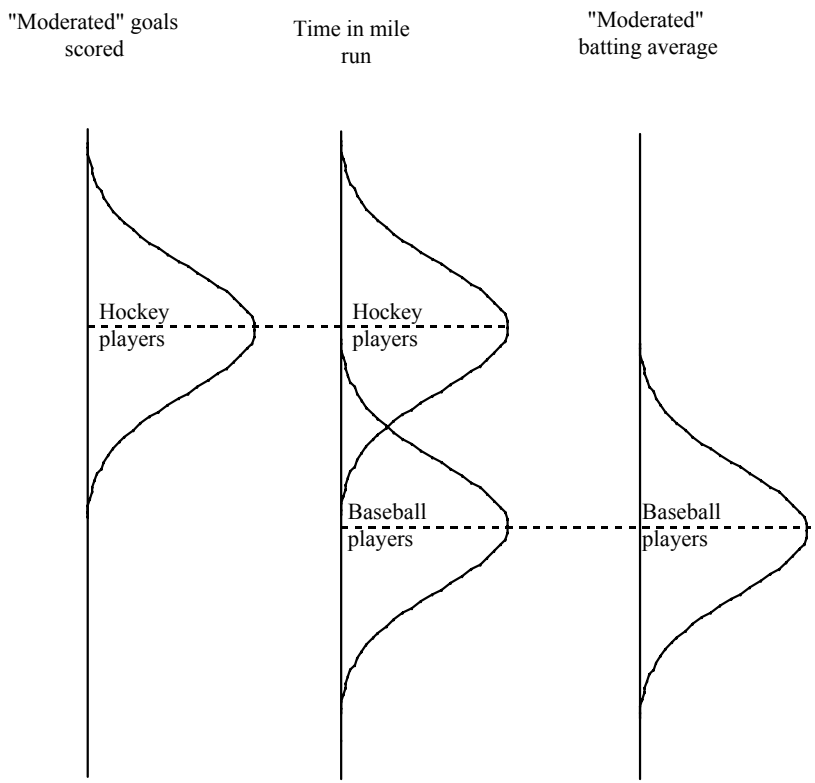

Figure 1

Moderated Distributions of Goals Scored and Batting Averages