Evidentiary Relationships Among Data-Gathering Methods and Reporting Scales in Surveys of Educational Achievement

CSE Technical Report 595

Robert J. Mislevy
CRESST/University of Maryland/
Educational Testing Service

April 2003

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.2 Validity of Interpretations and Reporting of Results
Robert Mislevy, Project Director, CRESST/University of Maryland/Educational Testing Service

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Institute of Education Sciences, or the U.S. Department of Education.

EVIDENTIARY RELATIONSHIPS AMONG DATA-GATHERING METHODS AND REPORTING SCALES IN SURVEYS OF EDUCATIONAL ACHIEVEMENT[1]

Robert J. Mislevy
CRESST/University of Maryland/Educational Testing Service

Abstract

Large-scale surveys of educational attainment gather data about the proficiencies of a sample of students to support inferences about the distribution in the populations. Several approaches to gathering data have been used, each with its own advantages and disadvantages. Several scales have also been used to bring results together from different students and different test forms. This paper lays out the evidentiary relationships between data gathered under five methods and inferences framed in terms of six reporting metrics. "Marketbasket reporting" receives special attention.

1.0 Introduction

The National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS) are examples of large-scale surveys of educational attainment. They assess knowledge and skills in samples of students, and then estimate the distribution of proficiency in the population and the relationships between proficiency and students' education and demographics. An assessment designer can choose from several data-gathering designs and several reporting scales. This paper examines the implications for analysis that result when different ways of gathering data are combined with different ways of reporting results. It highlights "marketbasket" reporting, which produces results in terms of a particular collection of tasks (Bock, 1993; Forsyth, Hambleton, Linn, Mislevy, & Yen, 1996; Johnson, 1996; Mislevy, 1998; National Research Council, 2000, 2001).

The main part of this presentation is organized as a matrix. The columns correspond to the methods by which one may collect student performance data. The rows correspond to reporting scales. The cells of the matrix lay out the relationship between data collected by a particular method and inferences desired on a particular

[1] Prepared for the NAS conference on marketbasket reporting in Washington DC, February 6-7, 2000. This paper is dedicated to Bert Green, who told me it was high time to sort this stuff out. For enlightening conversations and helpful suggestions, I thank Russell Almond, Paul Holland, Gene Johnson, John Mazzeo, David Rogosa, Neal Thomas, and my fellow members of the National Assessment Governing Board's Design/Feasibility Team: Robert Forsyth, Ronald Hambleton, Robert Linn, and Wendy Yen.

reporting scale, in terms of population and subpopulation averages and standard deviations, and proportions of the population above a predetermined criterion score (proportion above criterion, or PAC). Section 2 defines the data collection methods, inferential targets, and reporting scales. Section 3 describes machinery for probability-based reasoning from assessment data to target inference, namely, test equating and projection. Section 4 lays out the relationships between data and inferences in the cells of the table. The gist of the results appears in Table 1. Section 5 offers some comments on tradeoffs in assessment design. The overriding constraint: Designs ought to be restricted to combinations of data collection methods, reporting scales, targets of inference, and analytic procedures that produce sound inferences.

2.0 Background and Notation

This section defines the methods of data collection that correspond to the columns of Table 1, the reporting methods that correspond to its rows, and the targets of inference that are discussed in the cells.

2.1 Methods of Collecting Data

Methods for gathering assessment data differ as to cost, convenience, and implications for analysis. This section describes five basic approaches:

1. a single test form,
2. parallel forms,
3. tau-equivalent forms,
4. congeneric forms, and
5. unconstrained forms.

2.1.1 A Single Form

The simplest way to gather data is to administer an identical test form to every student from whom data are acquired. We'll denote the random variables that represent responses to the n items in this form, say Test X, by $X \equiv (X_1,...,X_n)$. Let $x \equiv (x_1,...,x_n)$ represent either realized values or generic values for a response vector. Let $S(x)$ be a function that maps a response vector to a summary score x^+. $S(x)$ could be defined simply as the count or the proportion of right answers, but variations include item weights, partial credit, "corrections for guessing," and item response theory (IRT) proficiency estimates. X^+ denotes the Test X total score as a random variable.

Table 1

Implications of Choices of Collecting Data and Reporting Results

	Method of collecting data				
Reporting scale	Single form	Parallel forms	Tau-equivalent forms	Congeneric forms	Unconstrained forms
Observed score on a particular administerable form (Marketbasket Reporting, Type 1)	[1,1] Piece of cake for all inferences! WYSIWYG	[1,2] Standard equating for all inferences.	[1,3] *Unequated* student-level scores work for means on percent-correct scale. Need 2-stage projections for *SD*s and PACs.	[1,4] Standard equating is generally unsatisfactory for means, worse for *SD*s and PACs. Need complex model: e.g., IRT and 2-stage projection.	[1,5] Need complex model: • test-level joint distribution, 1-stage empirical projections, or • MIRT, 2-stage projections.
True score on a particular administerable form (Marketbasket Reporting, Type 2)	[2,1] Means estimable by observed means. For other characteristics, need to estimate true-score distribution via strong true-score theory, or by IRT-based latent-distribution method then 1-stage projection back to MB true-score scale.	[2,2] MB means estimable by observed equated means. For other characteristics, must estimate true-score distribution via strong true-score theory, then equating function, or latent-distribution method then 1-stage projection to MB true-score scale.	[2,3] Means estimable by observed means. For other characteristics, need to estimate true-score distribution via strong true-score theory, or by IRT-based latent-distribution method, then 1-stage projection back to MB true-score scale.	[2,4] For all characteristics, need model-based (e.g., IRT, ARM) method to estimate distribution on some underlying scale, then 1-stage projection back to MB true-score scale.	[2,5] For all characteristics, need model-based method to estimate distribution on some underlying scale, then 1-stage projection to MB true-score scale. Differs from [2,4] in that may need multivariate underlying variable (e.g., MIRT).
True score on synthetic, non-administerable, form; e.g., domain score; very long representative form (Marketbasket Reporting, Type 3)	[3,1] If target reporting form is just a long version of the single administered form, then observed means estimate MB parameters but *SD*s and PACs don't. If not, need model to connect the single form with the hypothetical MB form. Then 1-stage projection from observed scores to the MB true-score distribution.	[3,2] For means, averages of observed scores have right expected values. For other characteristics, need more complex model: • MMS machinery (Lord, 1962; Sirotnik & Wellington, 1974). *Don't* add variance of observed score given examinee true score. • Estimate IRT-based latent-distribution, then use 1-stage projection back to true-score scale.	[3,3] For means, averages of observed scores have domain averages as their expected values. For other characteristics, need model-based latent-distribution method. then 1-stage projection to MB true-score scale.	[3,4] For all characteristics, need model-based method to estimate distribution on some underlying scale, then 1-stage projection back to MB true-score scale.	[3,5] For all characteristics, need model-based method to estimate distribution on some underlying scale, then 1-stage projection back to MB true-score scale. Differs from [3,4] in that may need multivariate underlying variable (e.g., MIRT).

(table continues)

Table 1 (continued)

	Method of collecting data				
Reporting scale	Single form	Parallel forms	Tau-equivalent forms	Congeneric forms	Unconstrained forms
Observed score on synthetic, non-administerable, form (Marketbasket Reporting, Type 4)	[4,1] If MB is just a long version of the single administered form, then observed means estimate targeted parameters; not so for *SD*s and PACs. In general, need model to connect the single form with the hypothetical target form. Then 2-stage projection from its observed scores to the MB observed-score distribution.	[4,2] For means, averages of observed scores have right expected values. For other characteristics, need more complex model: • MMS machinery (Lord, 1962; Sirotnik & Wellington, 1974). Do add variance of observed score given examinee true score. • Estimate IRT-based latent-distribution, then use 2-stage projection back to true-score scale.	[4,3] For all characteristics, latent-distribution method (e.g., IRT-based) then 2-stage projection to MB observed-score scale. For means only, averages of observed scores have MB averages as their expected values.	[4,4] Latent-distribution method (e.g., IRT-based), then 2-stage projection to MB observed-score scale.	[4,5] Latent-distribution method (e.g., IRT-based), then 2-stage projection to MB observed-score scale. Differs from [4,4] in that may need multivariate underlying variable (e.g., MIRT).
Examinee point estimates (MLE, Bayes, whatever—choose one!) on latent variable scale(s). *For cells [5,3]-[5,5], must specify a particular reference form*	[5,1] Point estimates from standard IRT program.	[5,2] Point estimates from standard IRT program. Variation among forms' information curves (ICs) introduces distortions. Less of a problem as ICs very similar.	[5,3] For all features, can use latent-distribution method (e.g., IRT-based), then 2-stage projection to reference-form point estimates. For means only, averages of point estimates observed scores have reference-form averages as expected values but only if long forms and no floor or ceiling effects.	[5,4] Use latent-distribution method (e.g., IRT-based), then 2-stage projection to observed-responses followed by mapping into designated point estimates for reference form.	[5,5] Latent-distribution method (e.g., MIRT-based), then 2-stage projection to observed-responses followed by point estimates for reference form. Reference-form scale can be based on a different latent variable model than the one used to model the responses.
Latent variable(s) on latent variable scale(s)	[6,1] IRT-based latent-distribution method. If long test or CAT so point estimates are stable, means of point estimates consistent with means of latent distributions. Not *SD*s or PACs, though.	[6,2] IRT-based latent-distribution method. If long test or CAT so point estimates are stable, means of point estimates consistent with means of latent distributions. Not *SD*s or PACs, though.	[6,3] Need model-based latent-distribution method.	[6,4] Need model-based latent-distribution method.	[6,5] Need model-based latent-distribution method. Differs from [6,4] in that may need multivariate underlying variable (e.g., MIRT).

In classical test theory (CTT; Gulliksen, 1950/1987; Lord & Novick, 1968), a student's observed test score, x^+, is conceived as the sum of two components: her "true score" t_x, plus an independent error term e_x that comes from a distribution with mean 0 and variance $\sigma^2_{e_X}$. That is,

$$x^+ = t_X + e_X.$$

A student's true score can be thought of as the expected value of her observed scores, or the average that would be obtained if she could somehow take the test many times with the same state of knowledge. CTT can be cast equally well in terms of total scores and percent-correct scores. It makes little difference when there is only one test form or a set of parallel tests, but we shall see that percents-correct have a useful property when working with tau-equivalent test forms.

We'll use **X** to refer collectively to the set of responses to Test X from a sample of students and $\mathbf{X}^+$ to refer to their scores.

2.1.2 Parallel Forms

A single form is easy to administer, but it usually can't cover a content domain well. We are interested in students' proficiencies across many skills, many concepts, and many relationships; it is difficult for a single test form to provide direct evidence on more than a relatively small sample of them. Either a single form is short enough to administer conveniently but doesn't cover the content domain, or it contains enough items to represent the domain but takes too long to administer. Collecting data with parallel test forms is one way to escape the dilemma. Parallel forms contain different items, but the same kinds and numbers of items, tapping the same mix of underlying skills. Different forms of the SAT, for example, are parallel.

Data gathered with parallel forms provide evidence about the same proficiency with the same accuracy. We will consider parallel forms that are all practical to administer as a standard way of gathering data. One of several parallel forms may be "special" when it comes to reporting, as "the" marketbasket form, but that is an issue of reporting and analysis rather than one of gathering data.

We will use the notation $Y \equiv (Y_1,\ldots,Y_n)$, $y \equiv (y_1,\ldots,y_n)$ $S(Y) \equiv Y^+$, and $S(y) \equiv y^+$ to denote the random variables and particular values of item responses and scores on a representative second test form, Test Y, that is parallel to Test X. **Y** and $\mathbf{Y}^+$ refer to the sets of responses and scores respectively of a sample of students to Test Y.

Under CTT, parallel tests have the same true score and independent errors that all have the same error variance. In terms of total scores or percents-correct, $x^+ = t_X + e_X$ and $y^+ = t_X + e_Y$, $\mu_{e_X} = \mu_{e_Y} = 0$, and $\sigma^2_{e_X} = \sigma^2_{e_Y}$.

2.1.3 Tau-Equivalent Forms

In classical test theory, "tau-equivalent" tests measure the same true score but may have different error variances (Novick & Lewis, 1967). In terms of percents-correct, tests with the same mix of items, but different numbers of them, are tau-equivalent. A student's expected percent-correct is the same in both cases, but a long form provides a more accurate estimate. One way of producing tau-equivalent forms is first to produce a set of parallel forms, then to administer different students tests constructed with various numbers of the parallel forms. For example, some students must take the items on three forms, whereas other students must respond to only one form.

We will use $Z \equiv (Z_1,...,Z_m)$, $z \equiv (z_1,...,z_m)$, $S(Z) \equiv Z^+$, and $S(z) \equiv z^+$ to denote variables associated with a representative Test Z, which is tau-equivalent to Test X.

Note that Test Z contains m items, while Test X contains n items. $\mathbf{Z}$ and $\mathbf{Z}^+$ are the responses and scores of a sample of students to Test Z. From the perspective of CTT, and with respect to percents-correct but not total scores, $x^+ = t_X + e_X$ and $z^+ = t_X + e_Z$ for any given student, and $\mu_{e_X} = \mu_{e_Z} = 0$, but $\sigma^2_{e_X} \neq \sigma^2_{e_Z}$.

Constructing tau-equivalent forms is only slightly less constraining than constructing parallel forms. We will see that tau-equivalent forms enable an assessor to gather data of different precision for different purposes (long forms for measuring individual students, short forms for estimating group averages or item statistics), yet report group averages in a common framework with familiar methods.

2.1.4 Congeneric Forms

"Congeneric" is another term from CTT that describes tests that "measure the same construct" but can have both different true-score variances and different error variances (Joreskog, 1971). We shall use the term to describe tests that include the same essential mix of knowledge and skills, but may differ as to the numbers, the difficulties, and the sensitivities of their constituent items. Constructing congeneric forms is less constraining than constructing tau-equivalent or parallel forms. As the examples below suggest, congeneric test forms can provide more information than common test forms because they can be better targeted to examinees.

We will use $W \equiv (W_1,...,W_m)$, $w \equiv (w_1,...,w_m)$, $S(W) \equiv W^+$, and $S(w) \equiv w^+$ to denote variables associated with a Test W that is congeneric to Test X. $\mathbf{W}$ and $\mathbf{W}^+$ are the responses and scores of a sample of students.

Although congeneric test forms are easier to construct than parallel or tau-equivalent forms, more complicated analyses are required to synthesize data across forms. Item response theory (IRT; Hambleton, 1989; Lord, 1980) is one way to do it. An IRT model expresses an examinee's tendency to perform well with respect to a domain of test items in terms of an unobservable proficiency variable. Each item *j* has parameters that characterize how responses to it depend on θ, through the IRT model $p(w_j|\theta)$. Item responses are assumed to be independent given θ, so the conditional probability of the response vector *w* is obtained as

$$p(w|\theta) = \prod_j p(w_j|\theta). \qquad \text{Eq. 2.1.4(1)}$$

Once a response vector *w* is observed, $p(w|\theta)$ is interpreted as a likelihood function. This is the basis of inference from *x* to θ and, in turn, from θ to other variables if required. For example, the model-based true score $t_W(\theta)$, defined as the expected score on Test W for an examinee with proficiency θ, is obtained as

$$t_W(\theta) = \int S(w)p(w|\theta)dw. \qquad \text{Eq. 2.1.4(2)}$$

In the special case of number-right scoring of right/wrong items, Equation 2.1.4(2) is the sum over items of their conditional probabilities of correct response:

$$t_W(\theta) = \sum_j p(w_j|\theta).$$

Most applications of IRT use a one-dimensional θ, but models with vector-valued θs are sometimes used (e.g., Reckase, 1997). These are multivariate IRT (MIRT) models. We will focus here on unidimensional IRT models, however: Tests are congeneric from the perspective of IRT if they are modeled in terms of the same unidimensional θ.

We find congeneric test forms in IRT-based computerized adaptive testing (CAT). Stocking and Swanson (1993), for example, show how to build tests that adapt their difficulties to each examinee's level of performance, yet maintain the same mix of item form and content.

The targeted test booklets in the 1984 NAEP Reading assessment (Beaton, 1987) are also congeneric test forms. All the booklets for the Age 9, Age 13, and Age 17 samples had similar mixes of reading comprehension items, some of which appeared in forms at all ages. The data from all these booklets were fit using the same IRT model. But many easy items appeared only in Age 9 booklets; others appeared in Age 9 and Age 13 booklets but not Age 17 booklets; some harder items appeared only in Age 13 and Age 17 booklets; and some appeared only in Age 17 booklets. In contrast, NAEP assessments such as Mathematics that measure several subscales and have different mixes of items from the subscales in different booklets are not using congeneric test forms.

2.1.5 Unconstrained Forms

"Unconstrained" forms consist of items from the same content domain but may differ, perhaps substantially, as to mix, number, format, and content. An example: A science assessment that administers to some students test forms with just multiple-choice questions across a range of subjects; to other students, forms based on hands-on experiments; and to still others, a series of tasks that all concern the same complex inquiry investigation.

Constructing unconstrained test forms imposes the fewest constraints on the assessment designer. Correspondingly, assessors who gather data with unconstrained test forms usually have to work hardest to bring the disparate results into a common reporting framework.

We will use $U \equiv (U_1,...,U_m)$, $u \equiv (u_1,...,u_m)$, $S(U) \equiv U^+$, and $S(u) \equiv u^+$ to denote variables associated with an arbitrarily composed Test U. $\mathbf{U}$ and $\mathbf{U}^+$ are the responses and scores of a sample of students.

2.2 Inferential Targets

"Choosing a reporting scale" means defining a variable in terms of which results will be expressed. The "targets of inferences" in a survey assessment are features of the distribution of that variable in some group of students. This presentation considers means, standard deviations, and proportions-above-criterion (PACs). The proportion of students above a NAEP Achievement Level and the percentage of candidates passing a certification examination with a fixed cut score are examples of PACs.

Let A be a random variable defined on the real numbers, and let $F(a)$ be the cumulative distribution function (cdf) in a population of interest; that is, $F(a) = \Pr(A \le a)$. We will denote the density function by $p(a)$, using the same notation whether A is continuous or discrete. In the continuous case, $F(a)$ and $p(a)$ are continuous functions; in the discrete case, $F(a)$ is a step function and $p(a)$ assigns probabilities to a countable set of points. Letting μ and σ denote the *mean* and *standard deviation* of A in this population,

$$\mu \equiv \int_{-\infty}^{\infty} a\, dF(a) = \int_{-\infty}^{\infty} a\, p(a) da$$

and

$$\sigma^2 \equiv \int_{-\infty}^{\infty} (\mu - a)^2\, dF(a) = \int_{-\infty}^{\infty} (\mu - a)^2\, p(a) da\,.$$

The proportion of the distribution above the fixed criterion value a_0 is obtained as $\mathrm{PAC}(a_0) = \int_{-\infty}^{a_0} dF(a)$. Equivalently,

$$PAC(a_0) \equiv \int_{-\infty}^{\infty} c(a;a_0) dF(a) = \int_{-\infty}^{\infty} c(a;a_0) p(a) da,$$

where $c(a;a_0)$ is an indicator function that takes the value 1 if $a \geq a_0$ and 0 if not.

Let G be a student background variable, concerning perhaps educational background or demographic characteristics. We may denote the cdf of A in the subpopulation in which G takes the value g by $F_g(a) = \Pr(A \leq a | G = g)$, and the density by either $p_g(a)$ or $p(a|g)$. Conditional means, standard deviations, and PACs for Group g are denoted as μ_g, σ_g, and $\mathrm{PAC_g}(a_0)$.

2.3 Reporting Scales

The preceding section discussed features of distributions that might be reported in survey assessments—but distributions of which variable? This section describes six possibilities:

1. observed scores on a particular administerable test form,
2. true scores on a particular administerable test form,
3. observed scores on a particular synthetic test form,
4. true scores on a particular synthetic test form,
5. point estimates on a latent variable scale, and
6. latent variable scales themselves.

We will refer to the first four approaches as types of marketbasket reporting. The fifth, point estimates on a latent variable scale, is generally not thought of as marketbasket reporting. We will see, however, that the notion of a reference test is usually required for it to provide sound inferences, even if that form is used explicitly for interpreting results.

2.3.1 Observed Score on a Particular Administerable Form (Marketbasket Reporting, Type 1)

In this reporting scale, a particular test form that can be conveniently administered to students, say Test X, is chosen. The items of Test X are considered a marketbasket of items, representative of the domain of interest. The *observed-score scale* of Test X, that is, concerning the distribution of X^+, is designated to be the reporting metric for any and all data collected. Whether or not data are actually collected using Test X, the information is expressed in terms of the X^+ scale.

This reporting metric has two attractive advantages. First, it is easy to interpret a score on this scale: It is simply the test score on a set of items we can see before us. Second, gathering data with this form leads directly to estimates of population distributions on the reporting scale. No transformations or adjustments of individuals' scores or school distributions are required, so if data are gathered with the marketbasket form itself, analyses are simple and can be carried out locally.

However, the limited sample of items that appears on such a form, and the equally limited scope of knowledge and skills it taps directly, limit the value of using a single form as the only way to gather data. As we shall see, though, data can be gathered in more flexible ways and the results mapped into the scale of a special marketbasket form, albeit through more complex analyses.

A second disadvantage is that the results are easy to interpret with respect to a given form because they are so tightly bound to that form. Consider a test of basketball players' ability to make free throws, and suppose we are interested in the proportion of players who can make more than 80% of their attempts. If we have a test that only consists of two attempts, we will have a much larger number of players whose observed success is above 80%—they made both shots—than if we take a larger sample, such as a hundred attempts. Classical test theory tells us that measurement error causes a consistent bias: The proportion of players whose observed percent of shots made is above 80% is an overestimate of the correct proportion whose true accuracy is above 80%, and the shorter the test the worse the overestimate tends to be. The reporting scale in the next section addresses this pitfall, but as we shall see, at the cost of greater complexity in analysis.

2.3.2 True Score on a Particular Administerable Form (Marketbasket Reporting, Type 2)

As above, a special test form that can be administered to students, Test X, is chosen. The *true-score scale* of Test X is now the reporting metric for any and all data collected. The items that constitute Test X are still the marketbasket of items, with the same advantages of interpretation and disadvantages of limited scope.

The distinction between Type 1 and Type 2 marketbasket reporting lies in the scale on which results are reported; there are no outward differences in the nature or constitution of the collection of items. With Type 1 marketbasket reporting, the actual observed score of student *i* on Test X, say x_i^+, is immediately on the desired reporting scale. With Type 2 marketbasket reporting, x_i^+ are viewed as random draw from a distribution whose expected value is the "true score" of Student *i,* say t_i; that is, $x_i^+ = t_{Xi}^+ + e_{Xi}$. True scores can be also be defined in terms of expected percents-correct rather than total scores with a simple rescaling. Either way, the Type 2 reporting scale is that of T_X rather than of X^+.

A key result from CTT is that an observed distribution of x^+s, no matter how large the sample of students, is generally not a good estimate of the distribution of T_X (Lord, 1969). Some model is needed to reason from any data to any inference on a true-score scale.[2] It is possible to estimate the proportion of basketball players whose accuracy is above 80% based on just two shots per player, but doing so requires knowing or assuming something about the shape of the distribution of their accuracies. Then the distribution of true accuracies can be estimated from the proportion of players making zero, one, and two of their attempts. Paradoxically, an accurate estimate of this true proportion be obtained if the number of players is large, even though just two shots per player provides a very unreliable estimate for each player individually.

A second key result is that estimated true scores have the same meaning across tau-equivalent forms (Section 4.2.3). This means that aside from sampling

[2] Different inferences lean on the model more heavily than others, though. Estimating the true-score mean of a population by its observed-score mean, for example, is less sensitive to alternative models than estimating the PAC for a high true-score cutoff.

variability, estimated true-score distributions are the same for every form in a set of tau-equivalent forms.

Under the assumptions of classical test theory (CTT), the error terms are independent across students, do not depend on their true scores, and have a common variance σ_e^2. A normal distribution is often assumed. Under strong true-score theory, more realistic distributions are posited for $p\left(x^+ | t_X\right)$, recognizing for example that X^+ may have to be an integer between 0 and n, or that the distribution of e_i may depend on the value of t_{Xi} (Lord, 1965; Lord & Novick, 1968).

2.3.3 True Score on a Synthetic Form (Marketbasket Reporting, Type 3)

A marketbasket of items that can be administrated conveniently represents only a fraction of the tasks that constitute a content domain. A larger collection of items would probe the domain more broadly and deeply, but would become too long to administer routinely, if at all. Bock (1993) has discussed advantages of content coverage and communication one gains by defining the reporting scale in terms of an entire item pool—the domain score reporting scale, in his terms (also see Bock, Thissen, & Zimowski, 1997). Alternatively, a shorter but proportionally representative synthetic form, again too large to administer conveniently but now better representing the skill domain than a short marketbasket form, could be defined and kept constant for reporting results while the actual item pool evolves.

We will refer to a synthetic form as Test V and suppose it contains q items. Let $\left(V_1,\ldots,V_q\right)$, $\left(v_1,\ldots,v_q\right)$, $S(V) \equiv V^+$, and $S(v) \equiv v^+$ denote the random variables and particular values of item responses and total scores on Test V.

Once a synthetic test form has been settled on, another choice must be made to determine the reporting scale. A Type 3 approach to marketbasket reporting is based on *true scores* for the synthetic form. The focus is then on the hypothetical distribution of expected scores of students, if they were to take the entire synthetic marketbasket form.

2.3.4 Observed Score on a Synthetic Form (Marketbasket Reporting, Type 4)

Again the assessor determines a synthetic test form, too long to administer conveniently but broadly representative of the domain. A Type 4 approach to marketbasket reporting is based on hypothetical observed scores for the synthetic form. How can one define a reporting scale in terms of observed scores for a test that will rarely, if ever, produce observed scores? Results using this approach have to be modeled projections from data on some other forms. As the length of a synthetic marketbasket form increases, though, the distinction between true scores and hypothetical observed scores vanishes.

2.3.5 Examinee Point Estimates on a Latent Variable Scale

A reporting scale for a survey of achievement can be defined in terms of IRT point estimates for individuals. As mentioned in Section 2.1.4, IRT gives the probability of an examinee's item responses in terms of an unobservable variable θ, say $p(u|\theta)$. IRT applications that concern the proficiency of individual examinees produce a point estimate from each $p(u_i|\theta)$ that is in some sense optimal for inference about the individual, along with a measure of its accuracy. The maximum likelihood estimate (MLE) $\hat{\theta}$ and its standard error $\sigma_{\hat{\theta}}$ are most common. The Bayesian posterior mean $\bar{\theta}$ and standard deviation $\sigma_{\theta|u}$ are also used, usually calculated with a common prior distribution $p(\theta)$ multiplying $p(u_i|\theta)$.

Defining a reporting scale in these terms means choosing a method of calculating point estimates and summarizing the evidence from each student's responses by her estimate. The underlying model can be either one-dimensional or multidimensional (MIRT). With MIRT, a function of the components of θ, such as a weighted average, can be used to summarize each student's results and serve as a composite reporting scale. The complexity of MIRT may be unavoidable when test forms differ substantially as to the mix of formats and skill demands their constituent items pose.

A key issue is that IRT point estimates generally have different distributions for different test forms. For any given θ, the distributions of $\hat{\theta}$s from repeated draws of x on Test X and w on Test W depend on the amount of information each test provides in the neighborhood of θ, as indicated by their test information curves. This depends mainly on the numbers of items in the tests and their difficulties relative to θ. The hypothetical distribution of $\hat{\theta}$s for repeated tests of an examinee with a given θ will be more dispersed as the form has fewer items in the neighborhood of θ. This means the distribution of a group of students' $\hat{\theta}$s from a short test will be wider than the distribution of their $\hat{\theta}$s from a long test, even if the IRT model is true. Distributions of point estimates from different test forms have comparable distributions only if the forms are closely parallel (technically, if their test information curves match).

Form-to-form vagaries in IRT point-estimate distributions don't preclude observed-score latent variable reporting with disparate test forms. The assessment designer can first designate some particular set of items and their associated parameters as a reference form. She can then project information obtained from various data-gathering forms into the observed-score scale of the reference form (Sections 3.5.4 and 3.5.5).

2.3.6 A Latent Variable Scale

Reporting with respect to a latent variable θ means synthesizing information from item responses in the form of an estimated distribution on the θ scale. Again θ can be unidimensional or multidimensional, and if it is multidimensional a function that produces a unidimensional summary can be defined. In all cases, though, this reporting option concerns the distribution of θ in the population of interest as opposed to the distribution of $\hat{\theta}$. Lord (1969) provides a pioneering exploration of this problem.

This is the reporting approach currently used in NAEP (see Mislevy, 1985, 1991, and Thomas, 1993, for the extension to the methods used in NAEP, which build on the methods of Rubin, 1987, for dealing with missing data in surveys). As we see in the following section, the same relationships and analytic machinery are also employed along the way to estimates of true- or observed-score estimates in some of the more ambitious combinations of data-gathering methods and reporting scales.

3.0 Tools for Reasoning From Data to Inferences

Under some combinations of data-gathering methods and reporting scales, the relationship between the resulting data and the desired inferences is quite straightforward. An obvious example is scores on a single form, meant to support inferences in terms of scores on that form. Other cells in the matrix are less intuitive. From the statistician's point of view, the issue is determining the probability distribution for the target inference that correctly expresses the evidence in the data. This section reviews two tools we need to accomplish this task in survey-type assessments, namely, equating and projection. Equating finds functions that map scores from two test forms onto one another, so that information from either can be used interchangeably. We will note that equating is possible only under highly constrained circumstances: Test forms must be parallel, or nearly so, and this happens only if great efforts have been made to construct them to achieve this result.

Projection is a generally applicable method, with commensurately more limited power. A projection maps what is known about one variable, in terms of a probability distribution, to what is known about another, also in terms of a probability distribution. Projection, as it is applied to linking assessments (Mislevy, 1993), incorporates results about posterior distributions and predictive distributions.

3.1 Equating

Parallel forms are constructed to provide essentially equivalent evidence across the range of inferences they are meant to support. They have the same kinds and numbers of items, the same mix of skill demands, the same timing and administration conditions, and the same scoring functions. Like Fahrenheit and Celsius temperature readings, though, data need not arrive on the same measurement scale. An equating function $E_{YX;P}\left(Y^{+}\right)$ maps Test Y scores to Test X scores with respect to a population *P* of students, with the intended result that

$$p\left(v \middle| E_{YX;P}\left(y^{+}\right)\right) = p\left(v \middle| y^{+}\right) \qquad \text{Eq. 3.1(1)}$$

for any variable *V* that is also defined for those students. That is, observing either y^{+} or the Text X score that the equating function maps y^{+} to, namely $E_{YX;P}\left(y^{+}\right)$, affects one's beliefs about *V* in exactly the same way.

Figure 1 depicts equating graphically. The link from the distribution of X scores to the distribution of Y scores is direct and point-by-point. The linkage depends on the marginal distributions of X scores and Y scores in Population P, but not directly on their joint distribution p(x,y). Note that the Test X axis is labeled with both *x* and $E_{YX;P}\left(y^{+}\right)$, since these both correspond to scores on Test X. The same reasoning holds for labeling the Y axis with both *y* and $E_{XY;P}\left(x^{+}\right)$, as they both correspond to scores on Test Y.

3.1.1 Bases of Equating Functions

When developers work hard to construct parallel forms and assessors administer and score them in the same way, the equating functions from one form to another all look very much like the identity function. Equating functions account for minor variations in difficulty or accuracy. There are many formulas and data collection designs for equating test forms (e.g., Kolen & Brennan, 1995), but all of them depend on Eq. 3.1(1) in some way. There are many features of distributions of Test X and equated Test Y distributions that would be identical if the forms were

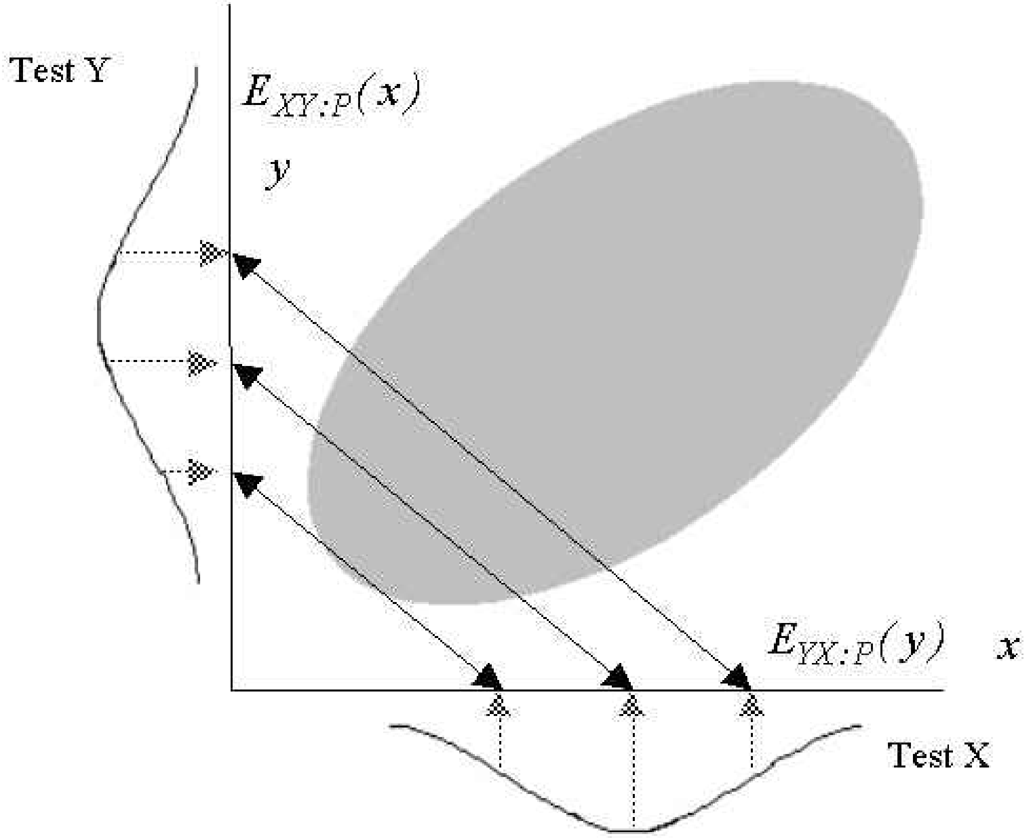


Figure 1. Equating Test X and Test Y with respect to Population P.

truly equated, and many features of relationships between Test X scores or equated Test Y scores with other variables that would be identical. An equating method selects certain ones of these features and finds the equating function that makes them match.

Some examples of equating: Equipercentile equating transforms Test Y scores so that in the equating sample of scores, their percentiles match up point by point. This implies that their means, standard deviations, and PACs match up. Linear equating transforms Test Y scores to make the transformed mean and standard deviation match those of Test X, but it doesn't change the shape of the distribution. These two methods are both examples of "observed-score equating," which means they match features of observed-score distributions. Such methods are by far most common in practice, and we will refer to them collectively as "standard equating." In contrast, "true-score equating" (Lord, 1980) makes expected test scores that correspond to the same true score match up. If the tests being equated are in fact nearly parallel, then results from any of the approaches will produce very similar equating functions.

3.1.2 Applying Equating Procedures to Tests That Are Not Parallel

Features of equated test distributions that are not built into an equating function need not match up, nor might features of their relationships with other variables. The degree to which they do anyway indicates just how similar the information from the test forms actually is (Dorans & Holland, 2000). When test forms differ substantially with respect to numbers of items, levels of difficulty, or mixes of knowledge and skill, equating functions that make some features match will fail on others.

An example of mismatches occurs when one attempts to equate tau-equivalent tests (Section 4.1.2 discusses its implications for analysis of assessment data). Consider a short form and a long form tapping the same skills, Test X and Test Z. Under CTT assumptions Examinee *i* has the same percent-correct true score on both tests: $x_i^+ = t_{Xi} + e_X$ and $z_i^+ = t_{Xi} + e_Z$. The error terms in both cases have zero expectation, but different variances, with $\sigma^2_{E_X} > \sigma^2_{E_Z}$. Since true scores are actually the same for all students in both tests, population and subpopulation means, standard deviations, and PACs are identical. For observed scores, population and subpopulation means are the same as each other and the same as their true-score counterparts as well: $\mu_X = \mu_Z = \mu_{T_X}$ and $\mu_{Xg} = \mu_{Yg} = \mu_{T_X g}$. But since $\sigma^2_X = \sigma^2_{T_X} + \sigma^2_{E_X}$ and $\sigma^2_Z = \sigma^2_{T_X} + \sigma^2_{E_Z}$, the observed-score variances are different, with $\sigma^2_X > \sigma^2_Z$.

It follows that PACs are generally different too. A larger proportion of students will have observed Test X scores above a high cutoff point x_0 than will have Test Z scores above the same point. Both overestimate the proportion whose true scores T_X are above that point. To illustrate, suppose both the population and errors are normal, and let $\Phi(z)$ denote the cumulative normal distribution. The expected proportions of values of T_X, X^+, and Z^+ above x_0 are given by $1-\Phi\left[\left(x_0-\mu_{T_x}\right)/\left(\sigma^2_{T_X}\right)\right]$, $1-\Phi\left[\left(x_0-\mu_{T_x}\right)/\left(\sigma^2_{T_X}+\sigma^2_{e_X}\right)\right]$, and $1-\Phi\left[\left(x_0-\mu_{T_x}\right)/\left(\sigma^2_{T_X}+\sigma^2_{e_Z}\right)\right]$. Suppose further that for percent-correct true scores T_X, u_{T_X} = .5 and $\sigma^2_{T_X}$ = .01, and for error variances, $\sigma^2_{e_X}$ = .01 and $\sigma^2_{e_Z}$ = .001. This implies reliabilities of .50 and .91 respectively for Text X and Test Z, values corresponding roughly to a 10-item test and a 100-item test. Now a cut score of .7 is two true-score standard deviations above the mean. The proportion of the population with true scores above the cut is about 2.3%, while the expected proportions of Test X and Test Y observed scores above the same cut point are 7.8% and 2.8% respectively. This situation is depicted in Figure 2.

True-score equating for Test X and Test Z in this situation gives the identity function by definition, since each student's expected percents-correct are the same for both tests. It follows that Test X and true-score equated Test Z population means and subpopulation means will match. But the variance of observed percents-correct for Test X is greater than that for Test Z, and as we have just seen, PACs don't agree either. Alternatively, an equipercentile equating for percent-correct scores matches up PACs, standard deviations, and means for the population as a whole—but does so by compressing the wider Test X. As a result, PACs, standard deviations, and means for subpopulations will not match up. With the numerical values of $\sigma^2_{e_X} = .5$ and $\sigma^2_{e_Z} = .1$, carrying out an equipercentile equating of Test Z to Test X matches total population means and variances to each other, but except for the population mean generally does not match them to the corresponding true-score values, and it does not match subpopulation means in the metric of equated scores. For example, if

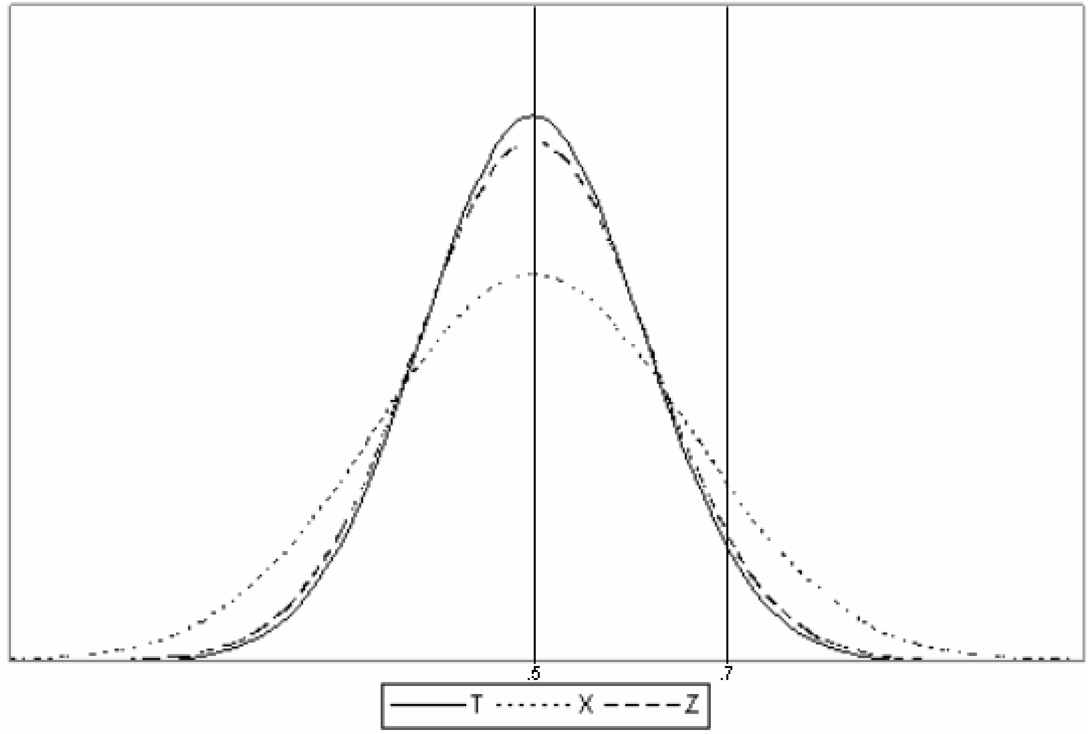


Figure 2. Differing proportions of percent-correct scores above cutpoint of .7, with respect to true scores T, a short test X, and a long test Z.

the mean of girls' true scores is 1.00, their expected mean on Test X is 1.00 but their expected mean on Test Z after equipercentile equating is .73

It is true nevertheless that certain "equatings" between nonparallel forms are routinely used in certain applied problems. They have proven over time to adequately support the particular inferences they are used for, with the particular kinds of data that are gathered.[3] Their success is better understood as an approximation of procedures that could be derived from first principles than as instructive applications of test equating. Other than an occasional passing comment, we will restrict the use of equating functions to parallel forms in this presentation.

3.2 Projection

A projection maps what is known about one variable, in terms of a probability distribution, to what is known about another, also in terms of a probability distribution. This section describes several kinds of projection that express the relationship between data from a given collection method to a given reporting scale. One-stage projections are discussed first; they project what is known about one variable to its implications for another. One-stage projections are further subdivided into those that move from an observed variable to a latent variable through a (mostly) theoretical model, and those that move from one observed variable to another through a (mostly) empirical model. Two-stage projections are then discussed. They move from an observed variable, to a latent variable, to a different observable variable, through mostly theoretical models in both steps. In each subsection, implications for both student-level and population-level inferences are considered.

3.2.1 One-Stage Model-Based Projection

For an individual student, one-stage projection from observations to a latent variable through a latent-variable model amounts to calculating the posterior distribution for that student's value on the latent variable. For a population, it might seem at first blush that one merely accumulates the individual-level results over the sample. We see, though, that calculating posteriors for individuals and estimating population distributions are parts of one inferential problem.

[3] An example is equating two parallel tests to each other by "equating" both to a third test that is not parallel to them. This works not because all three tests are equated in the sense we demand—equivalent evidence for any inference—but because for a given population the two parallel tests should have similar joint distributions with any third test that is correlated with them.

3.2.1.1 One-Stage Projection for Individual Students

Both CTT and IRT revolve around an expression for the probability of an observable variable conditional on an unobservable variable. In CTT, it is the probability of a test score given an examinee's true score,[4] for instance, $p\left(x^{+}|t_{x}\right)$ or $p\left(w^{+}|t_{w}\right)$. In IRT, it is the probability of an item response vector given the IRT proficiency variable, for instance, $p(x|\theta)$ or $p(w|\theta)$. As noted above, these expressions are interpreted as likelihood functions when the value of the observed variable is available. Belief about the latent variable is then obtained by Bayes theorem, in the form of the posterior density; for IRT,

$$p(\theta|x) = p(x|\theta)p(\theta)/p(x), \qquad \text{Eq. 3.2.1.1(1a)}$$

where $p(x) = \int p(x|\theta)p(\theta)d\theta$; and for CTT,

$$p\left(t_{x}|x^{+}\right) = p\left(x^{+}|t_{x}\right)p(t_{x})/p\left(x^{+}\right), \qquad \text{Eq. 3.2.1.1(1b)}$$

where $p\left(x^{+}\right) = \int p\left(x^{+}|t_{x}\right)p(t_{x})dt_{x}$.

Figure 3 illustrates one-stage projection. There is a joint distribution in Population P for the two variables *X* and θ. Prior to observing the value of *X* for say, Unit 1, belief about the value of θ for that unit is simply p(θ). If a value of X is observed for Unit 1, say x_1, then belief about the value of θ for Unit 1 is represented by $p(\theta|x_1)$. This distribution projects our knowledge about X into the space of θ. The expected value of this projection is $E(\theta|x_1)$. It is the average of the projected distribution for θ, but the spread of the distribution indicates that there remains uncertainty about the value of θ even after x_1 has been observed. Thus the mapping here is from a point to a distribution—unlike the case of equating, where mappings were one point to another one point between the variables being linked. Projection is not symmetric, in the sense that this term is used in the equating literature.

The introduction to Section 3.2 referred to the preceding formulas as a "mostly model-based" procedure. Indeed, the chain of reasoning runs directly through a model that relates the observed data to the latent variable. Different models may be checked against the data, however, and model parameters are generally estimated from data. A projection function depends on the relationship between the variables

[4] Strictly speaking, CTT addresses the first two moments of true and observed scores, not the full probability distribution. We use the extension to the full probability model in this presentation, assuming normal densities for both errors and population distributions.

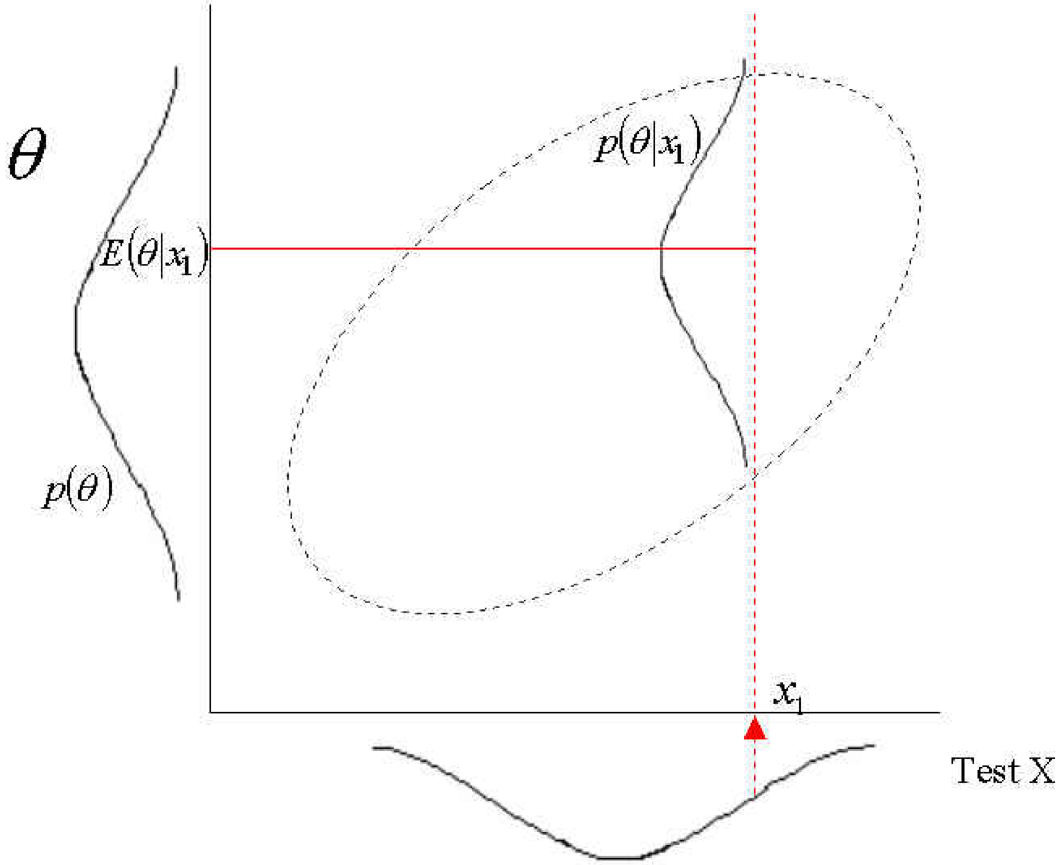


Figure 3. Projection from Test X to θ in Population P.

in a population of interest, so projection functions can be expected to vary from one population to another.

Projections can be used to draw inferences about an individual student in terms of the latent variable. Consider the question related to PAC reporting of whether Student *i* has a true score t_{Xi} above the cut-score t_{X0}. If t_{Xi} were known with certainty, we would simply compare t_{Xi} with t_{X0}. Using the notation $c(t_{Xi};t_{X0})$, the answer would be 1 if t_{Xi} is greater than or equal to t_{X0} and 0 if it isn't. However, if we know t_{Xi} only imperfectly, by having observed x_i^+, what we know is just a probability that Student *i* has a true score t_{Xi} above the cut-score t_{X0}:

$$\Pr\left(t_{Xi} \geq t_{X0} \middle| x_i^+\right) = \int c(t_X;t_{X0})p\left(t_X \middle| x_i^+\right)dt_X . \qquad \text{Eq. 3.2.1.1(2)}$$

Here $p\left(t_X \middle| x_i^+\right)$ is a probability density representing our belief about where t_{Xi} is likely to be, based on observing x_i^+; the probability that t_{Xi} is greater than t_{X0} is the proportion of that distribution that is above t_{X0}. In other words, the answer about whether t_{Xi} is greater than or equal to t_{X0} (1 if it is, 0 if isn't) is the weighted average of answers from all possible values of t_{Xi}, with each weighted by its probability given that x_i^+ has actually been observed.

Equation (3.2.1.1(1)) is readily extended to accommodate background information. For IRT,

$$p(\theta|x,g)=p(x|\theta,g)p(\theta|g)/p(x|g).$$

With rare exceptions,[5] the distribution of the observed variable given the latent proficiency variable is modeled as independent of background variables, so $p(x|\theta,g)=p(x|\theta)$ and

$$p(\theta|x,g)=p(x|\theta)p(\theta|g)/p(x|g), \qquad \text{Eq. 3.2.1.1(3a)}$$

with $p(x|g)=\int p(x|\theta)p(\theta|g)d\theta$. Similarly, for CTT,

$$p(x^+|t_x,g)=p(x^+|t_x)p(t_x|g)/p(x^+|g). \qquad \text{Eq. 3.2.1.1(3b)}$$

These subpopulation results have some surprising consequences, familiar in the psychometric literature since the 1920s as Kelley's paradox (Kelley, 1927). Consider a test being modeled under CTT, such that $X = \theta + e$ and e ~ N(0,10). Both Ann and Sam get scores of 50. Should we not have the same beliefs about their true scores? Suppose that for girls, θ ~ N(60,10), while for boys, θ ~ N(40,10). Standard Bayesian formulas for the posterior when both the prior and the likelihood are normal (e.g., Box & Taio, 1973) lead to posterior mean estimates for Ann and Sam of 55 and 45 respectively. Starting from identical likelihood functions, the posterior distribution for Ann is shifted upwards towards the girls' population distribution, while Sam's is shifted downward toward the boys' distribution. On the other hand, if we didn't happen to know that Ann and Sam were a girl and a boy respectively, the posterior distributions would still be the same—this time centered at the population average of 50. In the presence of measurement error and background information related to true scores, these regressed Kelley estimates are better estimates of students' true scores than are unbiased likelihood-based estimates that don't take subpopulation information into account, in the sense of having a smaller average squared error. There remain good reasons to use the unbiased estimates rather than the Kelley estimates when a test is being used "as a contest." We will see

[5] Such as differential item functioning (DIF), or performances on an item changing differently over time than the other items in a domain because of changes in curriculum or in society at large.

in next section, however, that the ideas behind Kelley estimates—namely, projections to true scores from observed scores—are necessary to recover unbiased subpopulation distributions.

3.2.1.2 One-Stage Projection for Populations

Note from Equation (3.2.1.1(3)) that projecting information about an examinee's latent variable given her responses involves the distribution that expresses belief before the responses are observed—for IRT, for example, $p(\theta)$ or $p(\theta|g)$, depending on whether background variables are involved. But in survey-type assessments, these distributions are typically not known. Indeed, estimating them is the goal of the survey in the first place. In these cases one draws inferences about the group and all sampled students simultaneously. Again using IRT as an example, we can approximate the density of θ from the response vectors **X=** $\{x_1,\ldots,x_N\}$ of a simple random sample of *N* students as

$$\begin{aligned} p(\theta)|\mathbf{X} &\approx N^{-1}\sum_{i=1}^{N} p(\theta|x_i) \\ &= N^{-1}\sum_{i=1}^{N} p(x_i|\theta)\,p(\theta)/p(x_i). \end{aligned} \qquad \text{Eq. 3.2.1.2(1)}$$

Note the appearance of $p(\theta)$ on both sides of the expression. In practice estimation procedures often presume a functional form for $p(\theta)$ and estimate its parameters. Several writers solved this problem in the 1970s and 1980s by maximizing a marginal likelihood function (e.g., Andersen & Madsen, 1977; Mislevy, 1984; Sanathanan & Blumenthal, 1978; more recently, see Cohen, 1997). This approach to estimating distributions of latent variables without using point estimates for each sampled individual is sometimes referred to as "direct estimates" of the population distributions (Cohen, 1997). With developments in computation and modeling, a full Bayesian approach (Gelman, Carlin, Stern, & Rubin, 1995) can now be brought to bear on the problem, as well as extensions to rater effects, covariate data, the hierarchical structure of classrooms, schools, districts, and states, and so on.

Once one does have a representation for $p(\theta)$ based on the observation of **X**, it is then possible to report on features of this distribution as they are required in reporting. PAC reports are particularly interesting because they raise the paradox introduced earlier with the free-throw accuracy assessment. Continuing with the IRT example from the preceding paragraph, suppose interest lies in $\text{PAC}(\theta_0)$, or the proportion of the population with θ values above θ_0. This proportion is approximated by the accumulation over students of the IRT equivalent of Equation 3.2.1.1(2):

$$\begin{aligned}\text{PAC}(\theta_0)|\mathbf{X} &\approx N^{-1}\sum_i \Pr(\theta \ge \theta_0 | x_i) \\ &= N^{-1}\sum_i \int c(\theta;\theta_0) p(\theta|x_i) d\theta.\end{aligned} \qquad \text{Eq. 3.2.1.2(2)}$$

The paradox is that when the result of Equation 3.2.1.2 is estimated from a sample of students, it does not generally agree with intuitive estimate obtained as the proportion of students whose point estimates are above the criterion, or $N^{-1}\sum c(\hat{\theta}_i;\theta_0)$. This result holds for Bayesian estimates, MLEs, or any other estimates that are optimal student-by-student. The size of the discrepancy depends on the amount of information about each student (more accurate tests lead to better agreement) and the extremeness of the criterion point (cutpoints near the middle of the distribution lead to better agreement than cutpoints in the tails).

If distributions for subpopulations or relationships with background variables are desired, the analogue of Equation Eq. 3.2.1.2(1) applies:

$$\begin{aligned}p(\theta|g)|\mathbf{X} &\approx N_g^{-1}\sum_{i=1}^{N_g} p(\theta|x_i, g) \\ &= N_g^{-1}\sum_{i=1}^{N_g} p(x_i|\theta) p(\theta|g) / p(x_i|g),\end{aligned} \qquad \text{Eq. 3.2.1.2(3)}$$

where N_g is the sample size for subpopulation *g* and the summation runs over only those students. It is important to note that when approximating $p(\theta|g)$, as it appears on the left side of the expression, it is necessary to also use $p(\theta|g)$ on the right side ("conditioning," in NAEP jargon; see Mislevy, 1991). Biases, sometimes serious, can result if one attempts to estimate $p(\theta|g)$ by first calculating posterior distributions for all sampled subjects with a single common prior distribution $p(\theta)$, then averaging over only those with *G*=*g* (Mislevy, Beaton, Kaplan, & Sheehan, 1992; Thomas, 2000).

3.2.2 One-Stage Empirical Projection

One-stage empirical projection moves from information in the form of one observable variable to another, through an empirically based joint distribution for the two of them. If a very large sample of subjects takes both Test X and Test U, we essentially know the joint distribution of test scores, $p(x^+, u^+)$, and thus the conditional distributions $p\left(x^+ \middle| u^+\right)$ and $p\left(u^+ \middle| x^+\right)$. (If we have a small sample of subjects who take both tests, we can estimate the joint distribution by smoothing or through a model, as in Rosenbaum and Thayer, 1987; this is what "mostly empirical" meant in Section 3.2.) If we observe a new student produce a score of u_0^+ on Test U, the predictive distribution that describes our belief about what his score on Test X would be is $p\left(x^+ \middle| u_0^+\right)$.

As with one-stage model-based projection, this procedure extends to subpopulations. We must work with the joint distribution for scores on both tests and the background variables as well, or $p\left(x^+, u^+, g\right)$, from which we obtain the required predictive distributions $p\left(x^+ \middle| u^+, g\right)$ and $p\left(u^+ \middle| x^+, g\right)$.

Figure 4 shows a simple bivariate distribution of discrete variables X^+ and U^+, in subpopulations of boys and girls. Each variable can take the values of 1, 2, 3, and 4. We will suppose these frequencies have been accumulated from many observations of both boys and girls, so they are known with great accuracy. What is the projection to U^+ if we observe Sam's Test X score is 3? We focus on the column where x^+=3 in the boys' table, which shows joint proportions of (.07, .10, .07, .04) respectively for U^+ values of 1, 2, 3, and 4. The projective distribution is obtained by normalizing these values, yielding (.25, .35, .25, .15). By similar calculation, we obtain a predictive distribution for U^+ if we observe Ann's Test X score is 3, except working through the girls' joint distribution table: (.18, .29, .29, .24). Note that this predictive distribution for U^+—initiated by the observation of the very same Test X score of 3—is flatter and shifted higher, due to the different shapes of the joint (X,Y) distributions among boys and girls.

Individual-level projections are the basis for projecting what is known about a population from observed data on one test form to a reporting scale based on observed scores of another. Suppose, for example, we have the predictive distribution $p\left(x^+ \middle| u^+, g\right)$ and observe the Test U scores $\mathbf{U}=\left\{u_1^+,\ldots,u_N^+\right\}$ of a sample of N_g students from subpopulation *g*. The projected distribution for Subpopulation *g* Test X observed scores is then obtained as

$$p\left(x^+ \middle| g\right)\mathbf{U} \approx N_g^{-1} \sum_{i=1}^{N_g} p\left(x^+ \middle| u_i^+, g\right)$$

$p(x^+,u^+ \mid \text{Sex=Boy})$

	x^+				
u^+	1	2	3	4	Sum
4	.01	.03	.04	.07	.16
3	.03	.04	.07	.10	.25
2	.04	.07	.10	.07	.30
1	.07	.09	.07	.04	.28
Sum	.16	.24	.30	.30	1.00

$p(x^+,u^+ \mid \text{Sex=Girl})$

	x^+				
u^+	1	2	3	4	Sum
4	.03	.04	.05	.07	.19
3	.04	.05	.07	.08	.24
2	.05	.07	.10	.08	.30
1	.07	.08	.08	.05	.28
Sum	.19	.24	.30	.28	1.00

Figure 4. Bivariate (x^+,u^+) distributions for boys and girls.

These ideas are readily extended to fitting a model for the full joint distribution of several test scores, conditional on G if inferences involving G will be desired. Then, given observation of any subset of scores, one can calculate the projection onto any function of the entire set. This includes a “marketbasket” summary score over some of the tests or all of them. The full joint distribution can be approximated even if data are not collected on the full set from any individuals, if we are willing to fit a simpler functional form. Beaton and Johnson (1990) applied this approach in the 1984 NAEP Writing assessment, calling it the “average response method,” or ARM. They fit a multivariate normal model for the joint distribution of 10 writing scores, with means modeled as functions of background variables G and a common residual covariance matrix. Assuming multivariate normality for the full joint distribution is a simplified functional form, which could be estimated from the matrix-sampled data collection design even though no student responded to more than three items. Beaton and Johnson could then calculate the projected distribution of the sum of an individual student’s hypothetical observed scores on all 10 items, conditional on the items he did take and the values of his background variables. From this they could project the distribution of his total score for the 10 items, which they used as the reporting scale (see Section 4.3.4).

3.3 Two-Stage Projection

Two-stage projection extends the ideas discussed above a step further: We are interested in what is known about one latent variable, say *X*, when observations arrive in the form of another, say *Y*, and the relationship between *X* and *Y* is expressed in terms of a third variable, say *Z*. In this presentation we will use two-stage projection to first project the distribution of a latent variable from an observed variable, and from the result calculate the predictive distribution of a different potentially observable variable. The effect is the same as in the one-stage empirical projection discussed above, also from one observable metric to another; the difference is adding a latent variable model link inside the chain of reasoning. Again, both individual-level and group-level inferences are of interest.

Projection across tau-equivalent forms through CTT serves to illustrate the form of two-stage projection. Tau-equivalent forms, Test X and Test Z, have the same mix of difficulties and skill demands, but they may vary in length. Under the assumptions of CTT, percent-correct observed scores X^+ and U^+ measure the same true score T_X:

$$X^+ = T_X + E_X \qquad \text{and} \qquad Z^+ = T_X + E_Z .$$

The error terms E_X and E_Z both have expectations of zero, but different standard deviations, σ^2_{eX} and σ^2_{eZ}.

If we observe a percent-correct score of x_i^+ from Student *i*, what is the projected distribution for his potentially observable z^+ score? The first stage of the projection is just as discussed in the preceding section: We use Bayes theorem to express the posterior distribution of the true score given x_i^+:

$$p\left(t_x \middle| x_i^+\right) = p\left(x_i^+ \middle| t_x\right) p(t_x) / p\left(x_i^+\right). \qquad \text{Eq. 3.3.1(1)}$$

Note again that the population distribution $p(t_x)$ is required here, and that if background data were involved, then $p(t_x|g_i)$ would appear instead. In the second stage, we calculate the predictive distribution of z^+ given what we have learned about t_x from x_i^+:

$$p\left(z^+ \middle| x_i^+\right) = \int p\left(z^+ \middle| t_x\right) p\left(t_x \middle| x_i^+\right) dt_x. \qquad \text{Eq. 3.3.1(2)}$$

Equation 3.3.1(2) can be interpreted as a weighted average of distributions of z^+ given t_x, each normal distributions with a mean of that particular t_x and a standard deviation of σ^2_{eZ}.[6] The weights for the various values of t_x are given by its posterior density given the less-then-certain evidence about t_x conveyed by x_i^+, or $p(t_x|x_i^+)$.

If we observe percent-correct scores of $\mathbf{X}^+ = \{x_1^+,...,x_N^+\}$ from N students, what is the projected distribution for the potentially observable z^+ scores of this sample? It is the accumulation of individual results like Equation 3.3.1(2):

$$p(z^+)|\mathbf{X}^+ \approx N^{-1}\sum_i p(z^+|x_i^+) = N^{-1}\sum_i \int p(z^+|t_X)p(t_X|x_i^+)dt_X. \qquad \text{Eq. 3.3.1(3)}$$

Suppose students are sampled randomly from a population and the underlying tau-equivalence model holds for the joint distribution of X^+, U^+, and T_X. Then for large samples of students, estimates of $p(z^+)$ based indirectly on sampled values of x^+ through Equation 3.3.1(3) will agree with estimates based directly on sampled values of z^+, even if test lengths are not long. This includes agreement on features such as means, standard deviations, and PACs.

Again, results that involve background variables must be calculated with projective distributions that incorporate those variables. The subpopulation counterpart of Equation 3.3.1(3), for example, is this:

$$\begin{aligned} p(z^+|g)|\mathbf{X}^+ &\approx N^{-1}\sum_i p(z^+|x_i^+, g) \\ &= N^{-1}\sum_i \int p(z^+|t_X)p(t_X|x_i^+, g)dt_X, \end{aligned} \qquad \text{Eq. 3.3.1(3)}$$

where $p(t_X|x_i^+, g) = p(x_i^+|t_X)p(t_X|g)/p(x_i^+|g)$. And again, failing to condition on subpopulation information generally gives an incorrect result.

4.0 The Matrix of Relationships Between Data Collection Methods and Reporting Scales

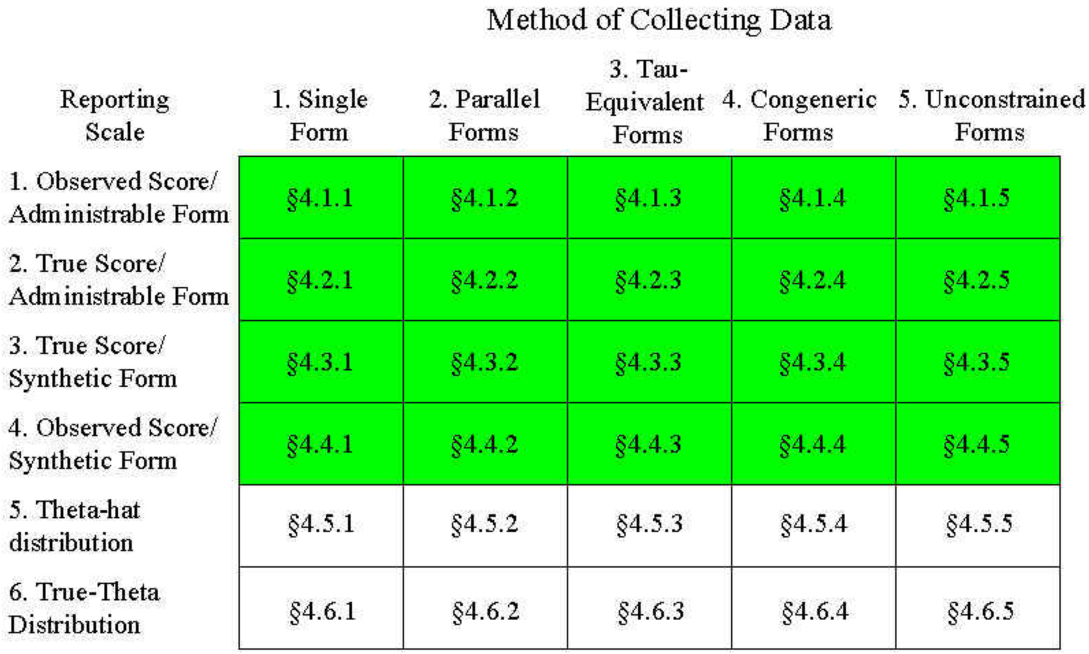
Section 2 defined the five methods of data collection and six reporting scales that constitute the columns and rows of a matrix, the entries of which will consider the relationship between data collected by a given method and inferences keyed to a given reporting scale. Section 3 reviewed the basic tools of probability-based reasoning we need in the discussion, and equating and projection in particular. This

[6] We obtain a closed form expression for this two-stage projection if the prior distribution and both of the likelihood functions are normal.

section puts the machinery to work. The following sections work across each row of the matrix, focusing on a particular reporting scale and discussing the implications for it that arise with each data collection method in turn. Figure 5 is a map of the discussions, in terms of sections of the paper. Most of the necessary ideas and techniques are required in early cells, so the cell-by-cell discussions become briefer as we proceed through the matrix.

Given a choice of a reporting scale (a row in the matrix), one can collect data in more than one way (multiple columns) and synthesize the evidence in terms of that same reporting scale. Working with the data in the different cells may require different analytic methods, though, each consistent with the evidentiary relationships entailed in each combination. This “single-reporting-scale/multiple-collection-methods” approach might be attractive for an assessment that serves multiple purposes, exploiting the advantages and disadvantages of different data collection methods for the chosen reporting scale.

Forsyth et al. (1996), for example, discuss using an observed-score scale for an administerable marketbasket form in order to capture some results easily and quickly (Cell [1,1]), albeit at the cost of domain coverage. An assessment system



	Method of Collecting Data				
Reporting Scale	1. Single Form	2. Parallel Forms	3. Tau-Equivalent Forms	4. Congeneric Forms	5. Unconstrained Forms
1. Observed Score/ Administrable Form	§4.1.1	§4.1.2	§4.1.3	§4.1.4	§4.1.5
2. True Score/ Administrable Form	§4.2.1	§4.2.2	§4.2.3	§4.2.4	§4.2.5
3. True Score/ Synthetic Form	§4.3.1	§4.3.2	§4.3.3	§4.3.4	§4.3.5
4. Observed Score/ Synthetic Form	§4.4.1	§4.4.2	§4.4.3	§4.4.4	§4.4.5
5. Theta-hat distribution	§4.5.1	§4.5.2	§4.5.3	§4.5.4	§4.5.5
6. True-Theta Distribution	§4.6.1	§4.6.2	§4.6.3	§4.6.4	§4.6.5

Figure 5. Map of discussions of relationships between reporting scales and data collection methods (shaded cells are marketbasket reporting approaches).

could make up on domain coverage by administering additional multiple forms that are congeneric to the marketbasket form or even constructed without constraints. The cost now is that reporting information from these forms on the desired scale involves some of the most complex relationships in the matrix (Cell [1,5]). As a compromise, these analyses could be carried out on a separate, more time-consuming but more comprehensive, stream of analyses some time after the initial results from the marketbasket data collection have been reported.

The following discussions focus on the structure of the evidentiary relationships between the data that are gathered and the inferences that are desired, not on the subtleties of complex student-sampling designs or the details of estimation procedures. To avoid these complications, we may think in terms of an extremely large simple random sample of *N* students in all cases.

4.1 Inference in Terms of Observed Scores on a Particular Administerable Form (Marketbasket Reporting, Type 1)

For this reporting scale, a particular administerable form, Test X, is designated to be a marketbasket of items. The *observed-score scale* of Test X concerns the distribution of scores X^+. This is the metric in which evidence collected by any method must be expressed. The cells in this row of the matrix concern different ways of collecting data to report on this scale. We shall see that the relationships between data collection and reporting scale become increasingly complex as we move to the right.

4.1.1 Data Collected Through a Single Form

This is the simplest cell in the entire matrix. It is based on what we might call "intuitive test theory." Any data that are gathered are gathered using the marketbasket test form itself, and observed scores on it are immediately on the desired reporting scale. The relationship between students' observed scores and population features can all be expressed as functions of scores that are calculated independently for each sampled student. In particular,

$$\mu_X \approx \bar{x}^+ = N^{-1}\sum_i x_i^+, \qquad \text{Eq. 4.1.1(1)}$$

$$\sigma_X^2 \approx N^{-1}\sum_i \left(x_i^+ - \bar{x}^+\right)^2, \qquad \text{Eq. 4.1.1(2)}$$

and, for the proportion of the population above the criterion value X_0^+, the proportion of students whose *observed scores* would be above X_0^+; i.e., $c(x_i^+; X_0^+)$=1 as opposed to 0:

$$\mathrm{PAC}(X_0^+) \approx N^{-1} \sum_i c(x_i^+; X_0^+). \qquad \text{Eq. 4.1.1(3)}$$

Analogues of these equations hold for subpopulations, the only difference being that the summations are only over the students in the subpopulation of interest:

$$\mu_{Xg} \approx \bar{x}_g^+ = N_g^{-1} \sum_{i:G=g} x_i^+, \qquad \text{Eq. 4.1.1(4)}$$

$$\sigma_{Xg}^2 \approx N_g^{-1} \sum_{i:G=g} (x_i^+ - \bar{x}_g^+)^2, \qquad \text{Eq. 4.1.1(5)}$$

and

$$\mathrm{PAC}_g(X_0^+) \approx N_g^{-1} \sum_{i:G=g} c(x_i^+; X_0^+). \qquad \text{Eq. 4.1.1(6)}$$

4.1.2 Data Collected Through Parallel Forms

This cell is slightly more complex than the preceding one, because at least some data are collected on a form other than the marketbasket form itself. There are two ways of approaching analysis in this cell. All forms are constructed to be parallel to the marketbasket form so we may either rely on test construction to give us essentially interchangeable scores across all forms, or use equating procedures to fine-tune the correspondence across forms. Under the former approach, we simply treat Test Y score as if it were a Test X score, and use the relationships outlined for Cell[1,1]. Under the latter approach, we thus replace each appearance of an x_i^+ in Equations 4.1.1(1)-(6) with a corresponding $E_{YX;P}(y^+)$. Data from several parallel forms can be combined in this way with one another and with data from Test X itself onto the Test X observed-score scale.

Of course even if we do carry out equatings, equating functions do not make information from parallel forms *exactly* interchangeable, so we can expect some uncertainty to enter the assessment system when we gather data with multiple forms that are only approximately parallel to the marketbasket form. The variance in the estimates of a parameter using data from different forms is properly part of the total uncertainty in its final estimate. Using fewer forms, or forms that are not actually very parallel, increases this component of error. One way to gauge this uncertainty, and then add it in to standard errors, is to "jackknife" estimates of population features with respect to data-gathering forms (Mosteller & Tukey, 1977). Fitting a hierarchical model that includes form-to-form variation is another way.

4.1.3 Data Collected Through Tau-Equivalent Forms

This combination contains a surprise: Standard equating a tau-equivalent form Test Z to the marketbasket form Test X makes inferences involving background variables worse rather than better. We will review why this is so, then discuss the projection relationships that hold in general to give correct results. Lord (1965) addressed the problem in this cell, with regard to lengthening a test and using a strong true-score theory model to project the characteristics of the new test's observed percent-correct score distribution.

4.1.3.1 The Perverse Consequences of Standard Equating in This Cell

Standard equating matches features of observed-score distributions. Recall from Section 3.1.2 that under the assumptions of CTT, tau-equivalent tests have the same true-score distribution on the percent-correct scale. Recall also that this implies the means of observed (percent-correct) scores for the population and subpopulations are also equal to one another and to their true-score counterparts as well; that is, $\mu_X = \mu_Z = \mu_{T_X}$ and $\mu_{Xg} = \mu_{Yg} = \mu_{T_X g}$. But since $\sigma_X^2 = \sigma_{T_X}^2 + \sigma_{E_X}^2$ and $\sigma_Z^2 = \sigma_{T_X}^2 + \sigma_{E_Z}^2$ while $\sigma_{E_X}^2 \neq \sigma_{E_Z}^2$, the observed-score variances and observed-score PACs do not agree. These relationships mean that observed-score population and subpopulation means $\bar{z}^+$ and $\bar{z}_g^+$ from a form that is tau-equivalent to the marketbasket *do* serve to estimate the correct (Type 1) marketbasket observed-score means, μ_X and μ_{Xg}, but other features of the Z^+ distribution *do not*.

In particular, their standard deviations do not agree—a (perceived) problem that standard equating was invented to solve. The linear equating function, for example, rescales Test Z scores by the factor that makes the standard deviations of the two forms match in the equating sample, namely $\sigma_{E_X} / \sigma_{E_Y}$. Applying an equipercentile or linear equating to Test Z scores to remedy the discrepancy in dispersion distorts the previously correct relationships among the means. Specifically, the mean of the equated *z*s for subpopulation *g*, or $N_g^{-1} \sum_g E_{ZX'P}\left(z_i^+\right)$, does not approximate the intended parameter μ_{Xg}, but rather the attenuated value $\left(\sigma_{E_X} / \sigma_{E_Y}\right)\mu_{Xg}$. Because the amount of attenuation depends on the relative sizes of the standard deviations, test length plays a major but wholly inappropriate role in determining the expectation of the estimated subpopulation mean. As a consequence, subpopulation standard deviations and PACs calculated from equated Test Z observed scores are also biased approximations of their Test X observed-score counterparts.

Summarizing the results of using standard equating in this cell, in an assessment with observed-score marketbasket reporting and data collected with forms tau-equivalent to the marketbasket form, inferences about percent-correct marketbasket means can be based on percent-correct means from any of the forms, *without standard observed-score equating*. Inferences about standard deviations or PACs from the same unequated subject-by-subject percent-correct scores are *not* correct, however. Applying standard equating appears at first to rectify problems with the population standard deviations and PACs, but in doing so distorts inferences concerning the relationships between proficiency and background variables.

4.1.3.2 Understanding the Relationship Through Projections

The relationship between observed scores on a marketbasket form and a different, nonparallel form can be expressed through projection. This may be either a one-stage empirical projection (Section 3.2.2) or a two-stage model-based projection (Section 3.3). The same relationship holds whether the other form is tau-equivalent, congeneric, or unconstrained, so this section lays out the formulas in tables in terms of the most general case, an unconstrained form Test U. The essential idea is to project belief about what the distribution of observed scores on Test X would be, given the observation of response vectors on Test U. We denote the projected means, variances, and PACs with double overscores.

The one-stage empirical projections in Table 2 address test scores rather than item responses. They presume a linking study or subsample that has provided the joint distribution of X^+ and U^+, for the population as a whole and conditional on any background variables *G* that will be involved in subsequent inferences. We denote these as $p\left(x^+,u^+\right)$ and $p\left(x^+,u^+\middle|g\right)$, from which we can obtain conditional distributions $p\left(x^+\middle|u^+\right)$ and $p\left(x^+\middle|u^+,g\right)$ (as described in Section 3.2.2).[7]

[7] Although the details of estimation are not a primary concern in this presentation, we point out that uncertainty about these conditional distributions should be taken into account when calculating the accuracy of final estimates. This is done in a full Bayesian solution which incorporates a higher level model for their parameters. In a multiple-imputation computational approximation, this amounts to

Table 2

Relationships Between Data Collected With an Unconstrained Form and Features of an Observed-Score Marketbasket Distribution: One-Stage Empirical Projection

Target of Inference	One-Stage Empirical Projection
Population Mean	$\overline{\overline{x^+}} = N^{-1}\sum_{i=1}^{N}\int x^+ p\left(x^+ \mid u_i^+\right)dx^+$
Subpopulation Mean	$\overline{\overline{x_g^+}} = N_g^{-1}\sum_{i=1}^{N_g}\int x^+ p\left(x^+ \mid u_i^+, g\right)dx^+$
Population Variance	$\overline{\overline{\sigma_X^2}} = N^{-1}\sum_{i=1}^{N}\int\left(x^+ - \overline{\overline{x^+}}\right)^2 p\left(x^+ \mid u_i^+\right)dx^+$
Subpopulation Variance	$\overline{\overline{\sigma_{Xg}^2}} = N_g^{-1}\sum_{i=1}^{N_g}\int\left(x^+ - \overline{\overline{x_g^+}}\right)^2 p\left(x^+ \mid u_i^+, g\right)dx^+$
Population $\mathrm{PAC}\left(X_0\right)$	$\overline{\overline{\mathrm{PAC}\left(X_0\right)}} = N^{-1}\sum_{i=1}^{N}\int c\left(x^+; X_0\right)p\left(x^+ \mid u_i^+\right)dx^+$
Subpopulation $\mathrm{PAC}\left(X_0\right)$	$\overline{\overline{\mathrm{PAC}_g\left(X_0\right)}} = N_g^{-1}\sum_{i=1}^{N_g}\int c\left(x^+; X_0\right)p\left(x^+ \mid u_i^+, g\right)dx^+$

A simple example of one-stage empirical projection from observations of u^+ to inferences about the distribution of x^+ can be generated from the $p\left(x^+, u^+ \mid g\right)$ distributions shown earlier as Figure 4. Figure 6 transforms the relationships into conditional distributions of x^+ given u^+ and g. Suppose that the observed data consist of only four observations, and they are u^+ values of 3 and 4 for two girls and 3 and 4 for two boys. Note that the boys' and girls' distributions are identical; both subpopulation samples, as well as the total population sample, have means of 3.5. Let us first work through the second row of Table 2, calculating the subpopulation average for girls. The projection to x^+ for the girl with a score of 3 is found in the u^+ = 3 row of the girls' conditional distribution in Figure 6, namely (.167, .208, .292, .333). The projection for the girl with a score of 4 is (.158, .211, .263, .368). The projection to

adding a layer of imputation from the posteriors of their parameters, as discussed in general terms in Rubin (1987) and particularized to NAEP-like applications in Thomas (1993, 2000).

$p(x^+,u^+|\text{Sex=Boy})$

	x^+				
u^+	1	2	3	4	Sum
4	0.067	0.200	0.267	0.467	1.000
3	0.125	0.167	0.292	0.417	1.000
2	0.143	0.250	0.357	0.250	1.000
1	0.259	0.333	0.259	0.148	1.000

$p(x^+,u^+|\text{Sex=Girl})$

	x^+				
u^+	1	2	3	4	Sum
4	0.158	0.211	0.263	0.368	1.000
3	0.167	0.208	0.292	0.333	1.000
2	0.167	0.233	0.333	0.267	1.000
1	0.250	0.286	0.286	0.179	1.000

Figure 6. Conditional distributions of x^+ given u^+ for boys and girls.

Test X for the girls' sample is the average of these, or (.162, .209, .277, .351), for x^+ scores 1 through 4 respectively. The projected x^+ mean for girls is therefore 2.82. By similar calculations using the conditional distributions for boys, we obtain the projected x^+ distribution of (.096, .183, .279, .442). Its mean is 3.07—a different value than obtained for girls, because the relationship between Test X scores and Test U scores is different for boys and girls. The overall projected mean is 2.94. Had we used a single projected distribution combining boys and girls, we would have recovered the correct projected total mean, but boys' and girls' x^+ means computed from the common projection would have (incorrectly) been identical.

A word on how these computations would be carried out using imputation methods. First, the empirical approximations of the conditional distributions $p\left(x^+ \middle| u^+, g\right)$ are calculated. Then, student by student, the appropriate projection to the x^+ metric is determined, in accordance with his or her values of u^+ and g. Then a value for x^+ is drawn at random from this distribution. A hypothetical sample of x^+ values is thus created for the entire sample of students. The expected values of this projected data set are correct for salient features of actual x^+ values, for the population as a whole and for subpopulations that have been conditioned on in the projection distributions.

The two-stage projections work through latent variable models. We presume a model in which the data that arise from both Test X and Test U depend on the same underlying proficiency variable θ. This latent variable may be either unidimensional, as in CTT and standard IRT, or vector-valued, as in multidimensional IRT (MIRT). Table 3 gives the equations for two-stage model-based projection based on test-score models such as CTT and strong true-score theory. Table 4 gives the equations for models at the level of response vectors, such as IRT and MIRT.

To offer some interpretation to the forms in these tables, let's take a closer look at the first row in Table 3, the test-level projected population mean for Text X scores, given observations in the form of congeneric Test U scores. The two stages of

Table 3

Relationships Between Data Collected With an Unconstrained Form and Features of an Observed-Score Marketbasket Distribution: Two-Stage, True-Score, Model-Based Projection

Target of Inference	Two-Stage Empirical Projection: True-Score Model (e.g., CTT)
Population Mean	$\overline{\overline{x^+}} = N^{-1}\sum_{i=1}^{N}\int x^+ \int p\left(x^+ \middle| t_X\right) p\left(t_X \middle| u_i^+\right) dt_X \, dx^+$
Subpopulation Mean	$\overline{\overline{x_g^+}} = N_g^{-1}\sum_{i=1}^{N_g}\int x^+ \int p\left(x^+ \middle| t_X\right) p\left(t_X \middle| u_i^+, g\right) dt_X \, dx$
Population Variance	$\overline{\overline{\sigma_X^2}} = N^{-1}\sum_{i=1}^{N}\int\left(x^+ - \overline{\overline{x^+}}\right)^2 \int p\left(x^+ \middle| t_X\right) p\left(t_X \middle| u_i^+\right) dt_X \, dx^+$
Subpopulation Variance	$\overline{\overline{\sigma_{Xg}^2}} = N_g^{-1}\sum_{i=1}^{N_g}\int\left(x^+ - \overline{\overline{x_g^+}}\right)^2 \int p\left(x^+ \middle| t_X\right) p\left(t_X \middle| u_i^+, g\right) dt_X \, dx^+$
Population $\text{PAC}\left(X_0\right)$	$\overline{\overline{\text{PAC}\left(X_0\right)}} = N^{-1}\sum_{i=1}^{N}\int c\left(x^+; X_0\right) \int p\left(x^+ \middle| t_X\right) p\left(t_X \middle| u_i^+\right) dt_X \, dx^+$
Subpopulation $\text{PAC}\left(X_0\right)$	$\overline{\overline{\text{PAC}_g\left(X_0\right)}} = N_g^{-1}\sum_{i=1}^{N_g}\int c\left(x^+; X_0\right) \int p\left(x^+ \middle| t_X\right) p\left(t_X \middle| u_i^+, g\right) dt_X \, dx^+$

Table 4

Relationships Between Data Collected With an Unconstrained Form and Features of an Observed-Score Marketbasket Distribution: Two-Stage Item-Level Model-Based Projection

Target of Inference	Two-Stage Model-Based Projection: Item-Level Model (e.g., IRT, MIRT)
Population Mean	$\overline{\overline{x^+}} = N^{-1}\sum_{i=1}^{N}\int S(x)\int p(x\|\theta)p(\theta\|u_i)d\theta\,dx$
Subpopulation Mean	$\overline{\overline{x_g^+}} = N_g^{-1}\sum_{i=1}^{N_g}\int S(x)\int p(x\|\theta)p(\theta\|u_i,g)d\theta\,dx$
Population Variance	$\overline{\overline{\sigma_X^2}} = N^{-1}\sum_{i=1}^{N}\int\int\left(S(x)-\overline{\overline{x^+}}\right)^2 p(x\|\theta)p(\theta\|u_i)d\theta\,dx$
Subpopulation Variance	$\overline{\overline{\sigma_{Xg}^2}} = N_g^{-1}\sum_{i=1}^{N_g}\int\left(S(x)-\overline{\overline{x_g^+}}\right)^2\int p(x\|\theta)p(\theta\|u_i,g)d\theta\,dx$
Population $\mathrm{PAC}(X_0)$	$\overline{\overline{\mathrm{PAC}(X_0)}} = N^{-1}\sum_{i=1}^{N}\int c(S(x);X_0)\int p(x\|\theta)p(\theta\|u_i)d\theta\,dx$
Subpopulation $\mathrm{PAC}(X_0)$	$\overline{\overline{\mathrm{PAC}_g(X_0)}} = N_g^{-1}\sum_{i=1}^{N_g}\int c(S(x);X_0)\int p(x\|\theta)p(\theta\|u_i,g)d\theta\,dx$

projection appear under the inner integral sign: $\int p(x^+|t_X)p(t_X|u_i^+)dt_X$, which condenses to $p(x^+|u_i^+)$, is the model-based projection of u_i^+ into the x^+ metric by way of the true score t_X. In the first stage of the projection, $p(t_X|u_i^+)$ uses Bayes theorem to express the probability distribution across all possible values of the true score what Student *i*'s value of it might be. Then $p(x^+|t_X)$ expresses the distribution of x^+ scores that would be expected from any given value of true score. The integral is the average over these latter x^+-predicted-from-t_X distributions, weighted by the projection into t_X from Student *i*'s observed u_i^+. The outer integral computes the average of the two-stage projected distribution for x^+, for Student *i*. The final summation, divided by the total number of students, calculates the average of these student-by-student projected x^+ expectations.

An imputation-based approximation of this formula highlights the two stages of projection. Again we start with $p\left(t_X \middle| u_i^+\right)$, the model-based projection from an observed Test U score to true scores. The first stage of projection is effected by drawing a value from it, say $\tilde{t}_{Xi}$. This is a plausible value for Student *i*'s true score, given the observed Test U score. Then the latent variable model provides a distribution for Test X scores given true scores, which we now instantiate with the true score fixed at $\tilde{t}_{Xi}$, or $p\left(x^+ \middle| \tilde{t}_{Xi}\right)$. The second stage of projection is effected by drawing a value of x^+ from this distribution, say $\tilde{x}_i^+$. This is a plausible value for Student *i*'s hypothetical Text X score, given the observed Test U score. The distribution of $\tilde{x}_i^+$ scores over students is a plausible representation of what their Test X score distribution might have been, had it rather than Test U scores been observed. In expectation, it yields the correct features of the Test X population distribution. If subpopulation features are desired, it is necessary to incorporate subpopulation membership into the first stage distributions in the projections, or $p\left(t_X \middle| u_i^+, g_i\right)$. Subpopulation membership is irrelevant in the second stage of the projection, however; by the conditional independence assumptions of latent variable models, $p\left(x^+ \middle| \tilde{t}_{Xi}, g_i\right) \equiv p\left(x^+ \middle| \tilde{t}_{Xi}\right)$.

The item-level two-stage projections summarized in Table 4 follow the same reasoning, although now projections involve response vectors of item responses to both Text U (observed) and Text X (projected). To prepare for the first stage of the projection, one estimates the distribution of the latent variable θ in the population from the entire collection of responses, again using methods such as Mislevy (1984) if inferences concerning subpopulation membership are not desired and Mislevy (1985) if they are. Then, the observed item response vector u_i for each Student *i* induces a likelihood function for the latent variable θ, which is combined with the appropriate population or subpopulation distribution via Bayes theorem to give a projected distribution for that student's θ. The latent variable model provides a distribution for Test X item responses via the latent variable model, $p(x \mid \theta)$. The distribution of a scoring statistic $S(x)$ for Student *i* can then be calculated with respect to this projection, and their distribution in a sample of students can be obtained by averaging the results over the sample.

The imputation approximation for item-level two-stage projection starts out like the score-level distribution, in that first the projection for the latent variable is calculated for each Student i via Bayes theorem given their observed response vector u_i and subpopulation data g_i if it is relevant. Then a value is drawn at random from this distribution, say $\tilde{\theta}_i$. Then a plausible vector of responses to the items of Test X for Student i is drawn from $p(x \mid \tilde{\theta}_i)$, say $\tilde{x}_i$. The desired score $S(\tilde{x}_i)$ is calculated. The distribution of such scores over students yields, in expectation, the correct distribution of $S(x)$ scores, had response vectors to Test X been observed.

4.1.4 Data Collected Through Congeneric Forms

Bringing information about the same underlying variable to the same observed-score reporting scale when test forms differ in difficulty is the so-called "vertical equating" problem. Although standard equating is often used for this situation to align standardized achievement tests at adjacent grades, some controversy surrounds the results and the practice itself. The consensus is that the results are best not considered "equated" even when the machinery of equating is used; they are better consider "scaled"—a weaker form of linking that may preserve unbiasness of estimates for individual students, but not features of population distributions (Kolen & Brennan, 1995, chap. 8). Congeneric forms are less similar to a marketbasket form than are tau-equivalent forms, so we can expect results from standard equating to be no better. Indeed, results are generally unsatisfactory for means, standard deviations, and PACs alike. The more complex relationships shown in Tables 2-4 must be applied instead.

4.1.5 Data Collected Through Unconstrained Forms

Tables 2-4 still apply when data are collected through unconstrained forms. Because the mix of skills and formats of tasks now varies across forms, however, more complex models are likely to be required to carry out model-based projection. For projection based on score-level models, for example, a factor analysis model might be required to account for the relationships among scores from different forms. For projection based on item-level models, MIRT might well be required rather than unidimensional IRT.

This is the way data are collected in the current NAEP. A five-dimensional IRT model was applied in the 1986 NAEP mathematics assessment, with the property that an item was presumed to depend on exactly one of the dimensions. The dimensions corresponded to subscales such as Algebra, Probability, and Numbers and Operations. This structure allowed test developers to vary the mix of items from the subscales across many test forms.

How would we project the distribution of observed scores for a marketbasket form in the current NAEP system? An 8-step procedure is listed below. It is one way to instantiate the equations in Table 4, using a multiple-imputation computing approximation. Steps 1-5 and Step 8 echo current NAEP analytic procedures (Cell [6,5] in Table 1, discussed in Section 4.6.5).

Step 1: Fit the IRT model to data from all forms, including the marketbasket form Test X and each regular booklet—we'll use Test U to denote a typical one of them—to estimate item parameters for all items.

Step 2: Estimate the conditional distribution of the five-dimensional θ conditional on background variables *G*, as described in, for example, Mislevy (1985) or Cohen (1997).

The multiple-imputations approximation requires creating *M* sets of pseudo-data. Steps 3-7 pertain to the creation of a single set, generically denoted set *m*.

Step 3: Draw a value for each parameter from the joint posterior distribution of the item parameters and the parameters of the conditional distributions. Treat these as known for Steps 4-7.

Step 4: For each sampled subject who responded to each test form U, calculate her posterior distribution for θ given her item responses and her background variables. Denote this as $p_m(\theta|u_i, g_i)$ for Student *i*.

Step 5: Draw a value from this distribution, and denote it $\tilde{\theta}_{im}$.

Step 6: For each item in the marketbasket form, draw a value from $p(x|\tilde{\theta}_{im})$, the predictive distribution of its response under the MIRT model, conditional on its item parameters and $\tilde{\theta}_{im}$. Call the resulting pseudo item-response vector $\tilde{x}_{im}$.

Step 7: Calculate the marketbasket form score $S(\tilde{x}_{im}) \equiv \tilde{x}_{im}^{+}$ for Student *i* in pseudo data set *m*.

Step 8: In the usual way of analyzing data from multiply imputed data (Rubin, 1987), calculate a feature of the distribution of X^{+}—statistics S(*X*) such as means and PACs—by calculating its value in each pseudo data set as if the $\tilde{x}_i^{+}$s were known with certainty, then averaging the results over pseudo data sets. We'll use the mean as an example. Each set *m* contains a plausible value for each Student *i*'s marketbasket score, say $\tilde{x}_{im}^{+}$. If the true mean is $\mu_{X^{+}}$, then from pseudo data set *m* we obtain the estimate

$$\overline{\overline{x^{+}}}^{(m)} = N^{-1}\sum_{i} \tilde{x}^{+}_{im}.$$

(Recall that we are assuming simple random sampling; weighted means may be required with complex samples.) The final estimate is the average of these within pseudo data set estimates:

$$\overline{\overline{x^{+}}}^{(m)} = M^{-1}\sum_{m} \overline{\overline{x^{+}}}.$$

In determining the precision associated with this estimate, there is a component due to sampling students and a component due to having observed values of *U* rather than values of *X*. Again assuming simple random sampling, the sampling variance associated with the estimate of the mean within each pseudo data set is the squared standard error of the mean, say $Var\left(\overline{\overline{x^{+}}}^{(m)}\right)$. The average of these values will be the contribution due to student sampling. The contribution due to uncertainty about *X* even from each sampled student is basically the variance among the different estimates of the mean from the *M* pseudo data sets. Altogether, then,

$$Var\left(\overline{\overline{x^{+}}} \middle| \mu_{X^{+}}\right) \approx M^{-1}\sum_{m} Var\left(\overline{\overline{x^{+}}}^{(m)}\right) + \left(\frac{M+1}{M}\right)\frac{\sum_{m}\left(\overline{\overline{x^{+}}} - \overline{\overline{x^{+}}}^{(m)}\right)^{2}}{M-1}.$$

For further information on creating and using multiple imputations in NAEP, the interested reader is referred to the NAEP Technical Reports (available from the National Center for Educational Statistics), Mislevy (1991), and Thomas (1993). For theory and more general background, the best reference is Rubin (1987).

An interesting feature of this cell in the reporting/data-gathering matrix is the balance of advantages and disadvantages it offers for different parties involved in an assessment. It is as simple as it can be for the user of reported data, since the interpretation is familiar and intuitively meaningful. It is as easy as can be for the constructor of test forms, since they don't need to maintain strict balances among forms as to length or content. It is as hard as it can be for the analyst, since it requires advanced methods in psychometrics, latent variable modeling, multiple imputation methodology, and, in complex designs, survey sampling variance estimation.

4.2 Inference in Terms of True Score on a Particular Administerable Form (Marketbasket Reporting, Type 2)

The similarity between this reporting scale and the preceding one is that in both, a particular administerable form, Test X, is designated to constitute a marketbasket of items. Scores on this set of tasks, actual or projected, are the terms in which performance is reported. But now the *true score* scale of Test X, that of expected values of Test X scores, is the metric in which evidence collected by whatever method must be expressed. Since by definition true scores are not observed for any students, it can be expected that the evidentiary relationships between observed data and reporting statistics will be more complicated.

Model-based inference is thus required under all data collection methods for this reporting scale, so let us recap the definitions of true scores under psychometric models. True scores can be defined either directly at the level of scores, as they are under CTT, or constructed from the level of items, as they are under IRT. Under CTT, by definition, a student's true score t_X on Test X is the expected value of her observed score X^+. Under the assumption of normality for errors, the distribution of the observed score for student *i* is $X_i^+ \sim N\left(t_{Xi}, \sigma_{e_X}\right)$. For a Test Y that is parallel to Test X, $Y_i^+ \sim N\left(t_{Xi}, \sigma_{e_X}\right)$ as well—a relationship that holds for either total scores or percent-correct scores. For a Test Z that is tau-equivalent to Test X, attention is focused on percent-correct scores, where the mean is the same but the error variance may differ, so $Z_i^+ \sim N\left(t_{Xi}, \sigma_{e_Z}\right)$

Under IRT, a student's true score for Test X in terms of either a total score or a percent-correct score is the expected value of that score, over all possible response patterns *x* to the items in Test X, with each pattern weighted by its probability given the student's θ, namely $p(x \mid \theta)$. Letting $S(x)$ be the desired scoring function, the model-based true score $t_X(\theta)$ is thus obtained as

$$t_X(\theta) = \int S(x)p(x|\theta)dx.$$

Recall that for number-right scoring of right/wrong items, $t_X(\theta)$ simplifies to the sum over items of their conditional probabilities of correct response.

4.2.1 Data Collected Through Single Form

The expected score of each student *i* is his or her true score, so it follows from the algebra of expectations that population and subpopulation means are the same for both observed scores and true scores. It is therefore possible to give estimates of *means* on the true-score scale of Test X without complex analyses. They are correctly estimated by their observed-score counterparts, and calculated by scoring each student's observed response vectors in the usual way and accumulating the results in the usual way. This holds for both total scores and percent-correct scores.

Although means match up neatly, other features of true-score and observed-score distributions do not; the distribution of observed scores x^+ for a population or subpopulation is not the same as the distribution of true scores t_X. In particular, the variance of the observed-score distribution is larger by the value of the error variance $\sigma^2_{e_X}$. This in turn overestimates the proportions of the population and subpopulations above high PAC cutoffs and below low cutoffs. As suggested in Figure 2, the bias is directly related to the magnitude of measurement associated with the test—its reliability, roughly—and increases as tests become less reliable. If features other than means are desired on the true-score scale of Test X, it is necessary to estimate the distribution of t_X by methods that lie beyond intuitive test theory—even when data are collected directly on Test X itself. This may be accomplished with CTT models at the level of test scores or IRT models at the level of item responses.

For test-level approaches, the key is having a model for an observed score given a true score, or $p(x^+|t_X)$. The CTT solution, in which it is posited that $x^+ = t_X + e_X$ with independent errors with the same variance $\sigma^2_{e_X}$ for all students, requires "deconvoluting" the observed-score distribution with respect to $\sigma^2_{e_X}$; that is, estimating what it would be if error terms with this variance were removed from all observed scores. Strong true-score theory (Lord, 1965; Lord & Novick, 1968, Part 6) has the same objective, but with more flexible and realistic error distributions. Table 5 gives the equations that map information from observed scores on Test X to projected features of the Test X true-score distribution. As before, it is necessary to estimate the true-score distribution, $p(t_X)$ or $p(t_X|g)$ as appropriate, in order to apply Bayes theorem and obtain the student-by-student projected true-score distributions $p(t_X|x) \propto p(x|t_X)p(t_X)$ or $p(t_X|x,g) \propto p(x|t_X)p(t_X,g)$. Features of these student-by-student projected true-score distributions are then accumulated to approximate features of population or subpopulation true-score distributions.

Table 5

Relationships Between Data Collected With Test X and Features of the Test X True-Score Distribution: One-Stage Model-Based Projection

Target of Inference	One-Stage Model-Based Projection
Population Mean	$\overline{\overline{t_X}} = N^{-1}\sum_{i=1}^{N}\int t_X p(t_X \mid x_i)dt_X$
Subpopulation Mean	$\overline{\overline{t_{Xg}}} = N_g^{-1}\sum_{i=1}^{N_g}\int t_X p(t_X \mid x_i, g)dt_X$
Population Variance	$\overline{\overline{\sigma_g^2}} = N^{-1}\sum_{i=1}^{N}\int (t_X - \overline{\overline{t_X}})^2 p(t_X \mid x_i)dt_X$
Subpopulation Variance	$\overline{\overline{\sigma_{Tg}^2}} = N_g^{-1}\sum_{i=1}^{N_g}\int (t_X - \overline{\overline{t_{Xg}}})^2 p(t_X \mid x_i, g)dt_X$
Population PAC(X_0)	$\overline{\overline{\mathrm{PAC}(t_{X0})}} = N^{-1}\sum_{i=1}^{N}\int c(t_X ; t_{X0})p(t_X \mid x_i)dt_X$
Subpopulation PAC(X_0)	$\overline{\overline{\mathrm{PAC}_g(t_{X0})}} = N_g^{-1}\sum_{i=1}^{N_g}\int c(t_X ; t_{X0})p(t_X \mid x_i, g)dt_X$

For item-level models, the key is having a model that gives the probability of a response vector to Test X as a function of a latent variable θ. IRT is the prototypical example of such a model. Estimating features of the true-score distribution of Test X from vectors of responses to Test X requires one-stage projection to the θ scale, then calculation of expected scores on Test X. Table 6 gives the equations that are required, in terms of the most general case of data collected from unconstrained forms. The relationships are the same for data collected on Test X itself, substituting x_i for each u_i. This cell is closest to the early investigations of Lord (1969).

4.2.2 Data Collected Through Parallel Forms

As with all methods of collecting data with the intent of reporting on the marketbasket true-score scale, an item-level latent variable model with one-stage projection provides for sound inference, through the relationships in Table 6.

Table 6

Relationships Between Data Collected With an Unconstrained Form and Features of a True-Score Marketbasket Distribution: One-Stage, Item-Level, Model-Based Projection

Target of Inference	One-Stage Model-Based Projection: Item-Level Model (e.g., IRT, MIRT)
Population Mean	$\overline{\overline{t_X}} = N^{-1}\sum_{i=1}^{N}\int t_X(\theta)p(\theta \mid u_i)d\theta$
Subpopulation Mean	$\overline{\overline{t_{Xg}}} = N_g^{-1}\sum_{i=1}^{N_g}\int t_X(\theta)p(\theta \mid u_i, g)d\theta$
Population Variance	$\overline{\overline{\sigma_T^2}} = N^{-1}\sum_{i=1}^{N}\int\left(t_X(\theta)-\overline{\overline{t_X}}\right)^2 p(\theta \mid u_i)d\theta$
Subpopulation Variance	$\overline{\overline{\sigma_{Tg}^2}} = N_g^{-1}\sum_{i=1}^{N_g}\int\left(t_X(\theta)-\overline{\overline{t_{Xg}}}\right)^2 p(\theta \mid u_i, g)d\theta$
Population PAC(X_0)	$\overline{\overline{\mathrm{PAC}(t_{X0})}} = N^{-1}\sum_{i=1}^{N}\int c(t_X(\theta); t_{X0})p(\theta \mid u_i)d\theta$
Subpopulation PAC(X_0)	$\overline{\overline{\mathrm{PAC}_g(t_{X0})}} = N_\mathrm{g}^{-1}\sum_{i=1}^{N_g}\int c(t_X(\theta); t_{X0})p(\theta \mid u_i, g)d\theta$

However, when data are collected with forms parallel to the marketbasket, certain inferences can be supported by simpler methods.

If Test Y is parallel to the marketbasket form Test X, then observed means of equated Test Y scores approximate true-score means of Test X. As with data from Test X itself, however, other features of equated Test Y scores do not approximate corresponding features of the Test X true-score distribution.

For distributional features other than means, one could estimate the true-score distribution for Test Y assuming CTT or strong true-score theory, then translate the resulting distribution through an equating function. Whether test forms are sufficiently parallel for this approach to provide satisfactory results is an empirical question. The best way to answer it is to apply the procedure with several parallel forms, and examine the variance across forms for estimated features in the uncertainty associated with the results. The more forms depart from parallelism, the greater this penalty term will be—perhaps even exceeding uncertainty due to sampling students, or the cross-time or subpopulation differences the survey was meant to estimate.

4.2.3 Data Collected Through Tau-Equivalent Forms

Again, an item-level latent variable model with one-stage projection correctly supports inference for reporting on the Test X marketbasket true-score scale, with data from a form Test Z that is tau-equivalent to Test X (Table 6).

When data are collected with forms tau-equivalent to the marketbasket, their means and subpopulation means on the percent-correct scale do approximate the corresponding Test X true-score counterparts; results for these statistics only are correct, with no further complications. Other features of the Test Z observed-score distribution do not generally approximate the corresponding Test X true-score distribution. Recall from Section 4.1.3 that standard equating destroys the correspondence between tau-equivalent forms with respect to population and subpopulation means. Only population variances and PACs come out correctly with population-level equating. Subpopulation means, variances, and PACs are biased, to a degree that depends on the difference in reliabilities of the two tests.

Because tau-equivalent forms have the same distribution on the true-score scale, CTT and strong true-score methods of estimating a true-score distribution of Test Z will immediately provide estimates of the true-score distribution of the marketbasket form Test X. The extent to which these results hold in practice depends on the extent to which forms are in fact tau-equivalent. This assumption can be checked empirically, and violations of it can be incorporated in statements of uncertainty for distributional features, for example by jackknifing with respect to test forms. That is, if tau-equivalence does not truly hold, then estimates of, say, subgroup means will differ more from data from different test forms than sampling error for student-selection alone would predict.

4.2.4 Data Collected Through Congeneric Forms

An item-level latent variable model with one-stage projection grounds inference for reporting on the Test X marketbasket true-score scale, with data from a form Test U that is congeneric to Test X (Table 6).

4.2.5 Data Collected Through Unconstrained Forms

An item-level latent variable model with one-stage projection grounds inference for reporting on the Test X marketbasket true-score scale, with data from an unconstrained form Test W from the same domain as Test X (Table 6), providing a satisfactory latent variable model is available that encompasses all items on both test forms. This cell differs from the preceding one only in that a more complex latent variable model, such as MIRT or the multiple IRT scales in current NAEP, is likely to be needed.

How would we project the distribution of true scores for a marketbasket form in the current NAEP system? The procedure very much like the one outlined in Section 4.1.5 for observed scores on a marketbasket form would work. Only Steps 6 and 7 need to change. There, Step 6 generated a random response vector $\tilde{x}_i$ for the marketbasket form for Student *i*, drawn from $p\left(x\middle|\tilde{\theta}_i\right)$, and Step 7 evaluated the score for this vector, $\tilde{x}_i^+ \equiv S(\tilde{x}_i)$. Here, we replace $\tilde{x}_i^+$ with the Test X true score that corresponds to $\tilde{\theta}_i$, or $t_X\left(\tilde{\theta}\right)_i$. For a test comprised of dichotomous items, this is the total or the average, as required, of the probabilities of correct response to each Test X item as predicted though the IRT model, using each student's projected θ distribution in turn. We may then study distributions of these imputations—features of their distributions within pseudo data sets, in light of variation of those features across pseudo data sets.

4.3 Inference in Terms of True Score on a Synthetic Form (Marketbasket Reporting, Type 3)

This section concerns reporting on the true-score scale of a synthetic marketbasket Test V, a specified collection of items so long that it will probably never be administered to students. We will denote the true-score variable and a particular value of it by T_V and t_V respectively, and, when using a latent variable model, refer to the true score associated with a particular value of θ as $t_V(\theta)$.

For all methods of data collection—that is, for each cell in this row—the item-level model projections described in the preceding section on true-score reporting for an administerable marketbasket form apply to a synthetic marketbasket form. The relationships are shown in Table 7. Only relatively minor changes in notation have been required from Table 6, namely replacing every appearance of t_X and any variation of it with t_V or the corresponding variation of it.

Table 7

Relationships Between Data Collected With an Unconstrained Form and Features of a True-Score Synthetic Marketbasket Distribution: One-Stage, Item-Level, Model-Based Projection

Target of Inference	One-Stage Model-Based Projection: Item-Level Model (e.g., IRT, MIRT)
Population Mean	$\overline{\overline{t_V}} = N^{-1} \sum_{i=1}^{N} \int t_V(\theta) p(\theta \mid u_i) d\theta$
Subpopulation Mean	$\overline{\overline{t_{Vg}}} = N_g^{-1} \sum_{i=1}^{N_g} \int t_V(\theta) p(\theta \mid u_i, g) d\theta$
Population Variance	$\overline{\overline{\sigma_V^2}} = N^{-1} \sum_{i=1}^{N} \int \left(t_V(\theta) - \overline{\overline{t_V}} \right)^2 p(\theta \mid u_i) d\theta$
Subpopulation Variance	$\overline{\overline{\sigma_{Vg}^2}} = N_g^{-1} \sum_{i=1}^{N_g} \int \left(t_V(\theta) - \overline{\overline{t_{Vg}}} \right)^2 p(\theta \mid u_i, g) d\theta$
Population $\mathrm{PAC}(X_0)$	$\overline{\overline{\mathrm{PAC}(t_{V0})}} = N^{-1} \sum_{i=1}^{N} \int c(t_V(\theta); t_{V0}) p(\theta \mid u_i) d\theta$
Subpopulation $\mathrm{PAC}(X_0)$	$\overline{\overline{\mathrm{PAC}_g(t_{V0})}} = N_{\mathrm{g}}^{-1} \sum_{i=1}^{N_g} \int c(t_v(\theta); t_{V0}) p(\theta \mid u_i, g) d\theta$

The general relationships having thus been covered, the following subsections comment on additional considerations that arise under various data collection methods.

4.3.1 Data Collected Through Single Form

Suppose data are collected from all examinees with the single form Test X. If Test V is just a longer version of Test X, then Test X is tau-equivalent to Test V. It follows that observed Test X means for populations and subpopulations approximate their synthetic marketbasket counterparts on the marketbasket true-score scale. Other features of the distribution, including PACs, do not.

How might such a design come about, with data collected with only one test form but results reported in terms of true scores on a much larger form? One way might be for a school district to administer a survey of items sampled from a published item bank that includes IRT parameters. The results from the single form could then be projected to results in the item bank—specifically, to estimate the proportion of items in of the item bank that students would answer correctly. Projection would be required to obtain the correct population and subpopulation means, however; student-by-student point estimates on the IRT scale would yield biased estimates of even population and subpopulation means, when substituted into the IRT model and used to calculated expected percents-correct item-by-item. The size of the biases is not mitigated by large samples of students, but they do decrease as test length increases.

4.3.2 Data Collected Through Parallel Forms

Suppose data are collected using a number of test forms that are (a) parallel to one another and are administered and (b) tau-equivalent to the synthetic marketbasket. In this case population and subpopulation averages of observed percent-correct scores approximate the corresponding means of the synthetic test true-score scale. This relationship does not hold unless the strong assumption of tau-equivalence is satisfied, and even when it is, features of the distributions other than overall and subpopulation means do not generally match the features of the marketbasket true-score percent-correct distribution.

One way of guaranteeing tau-equivalence is to sample items for parallel tests from a well-defined item pool, stratifying with respect to important characteristics of skill and format. The pool as a whole constitutes the synthetic form, Test V.[8] When this is done, it is possible with more complex analytic machinery to estimate not only means but higher moments of the Test V distribution. The machinery was developed in the 1960s and 1970s by researchers including Lord (1962) and Sirotnik and Wellington (1974), under the rubric of multiple matrix sampling (MMS). When estimating distributions of Test V *true* scores by these methods, variation of observed scores around expected scores is not added in to estimates of features of the population distribution.

4.3.3 Data Collected Through Tau-Equivalent Forms

Suppose data are collected using a number of test forms that are (a) tau-equivalent to one another and are administered and (b) tau-equivalent to the synthetic marketbasket. Then population and subpopulation averages of observed

[8] This is effectively the design used in the California Assessment Program (CAP) in the late 1970s and early 1980s (Carlson, 1979).

scores approximate the corresponding means of the synthetic test true-score scale. This relationship does not hold unless the assumption of tau-equivalence is satisfied, and even when it is, features of the distributions other than overall and subpopulation means do not generally match the features of the marketbasket true-score percent-correct distribution.

4.3.4 Data Collected Through Congeneric Forms

In general, one needs a latent variable model and the relationships from Table 7 to bring data collected in this manner to the synthetic marketbasket true-score scale. An interesting application in this cell is the ARM (Average Response Method) reporting approach that Beaton and Johnson (1990) developed for reporting performance in the 1984 and 1986 NAEP assessments of writing. The writing domain was defined by about 20 tasks, each scored on a 0-4 integer scale. A given sampled student would be administered only between 1 and 3 tasks. Beaton and Johnson used missing-data methods to estimate the joint distribution of all tasks, in terms of multivariate normal distributions conditional on grade and age, gender, ethnicity, and other key reporting variables. A custom-built linear regression model, tailored to the tasks an individual student was administered, was then used to project the expected score on the entire domain.

4.3.5 Data Collected Through Unconstrained Forms

As for the preceding cell, one needs a latent variable model and the relationships in Table 7 to bring data collected in this manner to the synthetic marketbasket true-score scale. It differs only in that a more complex latent variable model is probably needed. Reporting of this type could be incorporated into the current NAEP system by the methods discussed in Section 4.2.5, which itself was a modification of the scheme more fully described in Section 4.1.5.

4.4 Inference in Terms of Observed Score on a Synthetic Form (Marketbasket Reporting, Type 4)

This section concerns reporting on the observed-score scale of a synthetic marketbasket Test V. If the marketbasket is very long, as it would certainly be with the domain-referenced scores Bock (1993) and his colleagues favor, there is no practical difference between observed scores and true scores on the marketbasket form.

Bringing evidence from all the forms of data collection to this scale can be accomplished with two-stage projection through an item-level latent variable model. That is, the items on the administered form, say Test U, to infer distributions in terms of the latent variable θ; these distributions are used in turn to infer distributions on the potentially observable items that constitute Test V. The essential relationships in Table 4, which were developed for projection to the observed-score scale of an administerable marketbasket form, apply. One need only replace each appearance of x or some variation of it with v or the corresponding variation.

The generally applicable relationships having been thus described, the following sections offer additional comments about various cells in this row.

4.4.1 Data Collected Through Single Form

If data are collected from all examinees with a single form Test X that is a representatively shortened version of the synthetic marketbasket form Test V, then Test X is tau-equivalent to Test V. Observed Test X means for populations and subpopulations approximate their synthetic marketbasket counterparts on the marketbasket observed-score scale for percents-correct. Other features of the distribution, including PACs, do not.

4.4.2 Data Collected Through Parallel Forms

If a number of parallel forms are administered and all are tau-equivalent to the synthetic marketbasket, then population and subpopulation averages of observed scores approximate the corresponding means of the synthetic test percent-correct observed-score scale. This relationship does not hold unless the strong assumption of tau-equivalence is satisfied, and even when it is, other features of the distributions do not match those of the marketbasket observed-score distribution.

Suppose the tau-equivalent forms are constructed by sampling items to create parallel tests from a well-defined item pool, and the pool as a whole constitutes the synthetic form, Test V. The methods of multiple matrix sampling (Lord, 1962; Sirotnik & Wellington, 1974) can again be used to estimate means and higher moments of the Test V distribution. When estimating distributions of Test V *observed* scores by these methods, the component due to variation of observed scores around expected scores *is* added in to estimates of features of the population distribution.

4.4.3 Data Collected Through Tau-Equivalent Forms

If a number of tau-equivalent forms are administered and all are tau-equivalent to the synthetic marketbasket, then population and subpopulation averages of observed percent-correct scores approximate the corresponding means of the synthetic test percent-correct observed-score scale. Even when the assumption of tau-equivalence holds, though, other features of the distributions do not match.

4.4.4 Data Collected Through Congeneric Forms

In general, one needs a latent variable model and the two-stage projection relationships from Table 4 (again revised to reflect Test V rather than Test X as the reporting scale) to bring data collected in this manner to the synthetic marketbasket observed-score scale. The idea is the same as in that cell: One begins by using observed data on Test U to estimate the distribution of a latent variable θ, conditional on whatever background variables are of interest. In the first stage of the two-stage projection, predictive distributions of θ are created for each student given u and g. In the second stage, a projected distribution of the statistic $S(v)$ is obtained with respect to the predicted θ distribution. When imputation approximations are used, the first stage for each student amounts to drawing a value $\tilde{\theta}_i$ from $p(\theta|u_i, g_i)$, then drawing values for item responses to each of the items in the synthetic marketbasket Test V, from $p(v|\tilde{\theta}_i)$. These (possibly very long!) hypothetical response vectors are used to calculate the values of the statistics of interest $S(v_i)$ student by student.

4.4.5 Data Collected Through Unconstrained Forms

As for the preceding cell, one needs a latent variable model and the notationally modified relationships from Table 4 to bring data collected in this manner to the synthetic marketbasket observed-score scale. It differs only in that a more complex latent variable model is probably needed. Reporting of this type could be incorporated into the current NAEP system by the methods discussed in Section 4.1.5.

4.5 Inference in Terms of Point Estimates on a Latent Variable Scale

A reporting scale for a survey of achievement can be defined in terms of point estimates for individuals' latent variables. Under CTT, where unbiased estimates of individuals' true scores are simply their observed scores, this approach simplifies to reporting on the observed-score scale of the marketbasket—the first row of the matrix, which has already been discussed. Therefore, in this section we will focus on latent variable models defined at the level of items, namely IRT models. An IRT model gives the probability of a response vector u in terms of an unobservable variable θ, or $p(u|\theta)$. Once a response vector u is observed, $p(u|\theta)$ is interpreted as a likelihood function. In applications that focus on individual students, a point estimate that is in some way optimal for measuring individuals is calculated. The maximum likelihood estimate (MLE) $\hat{\theta}$ is the point at which $p(u|\theta)$ has the highest value. The Bayesian estimate $\bar{\theta}$ is the mean of the posterior distribution, usually calculated with a common prior distribution $p(\theta)$ multiplying $p(u|\theta)$. Alternatively, Bayesian point estimates can be calculated taking a student's background variables into account as well, by applying $p(u|\theta, g)$ rather than $p(u|\theta)$. Note that to carry this out correctly in either case requires estimating the latent variable distributions $p(u|\theta)$ or $p(u|\theta, g)$ by means of one of the direct estimation procedures mentioned earlier.

One may consider defining a reporting scale for a survey by designating a method of calculating point estimates and summarizing the evidence from each student's responses in its terms. But IRT point estimates generally have different distributions for different test forms, depending on the numbers of items in the tests and their difficulties relative to any given θ. For example, the distribution of a group of students' $\widehat{\theta}$ s from a short test will be wider than the distribution of their $\widehat{\theta}$ s from a long test, even if the IRT model is true. Except in the special cases discussed in the following subsections, attempting to define a reporting scale in terms of IRT point estimates without designating a particular reference form or parallel set of reference forms is likely to fail. These kinds of problems make reporting on an IRT point-estimate scale a hazardous matter when only student-by-student point estimates are the basis of inference.

This row of the matrix represents a trap that more than one large-scale assessment program has fallen into: IRT student-level point estimates have been used as the basis of reporting scores in a base year, and it is desired to report comparable scores in a subsequent year when data are gathered with revised test forms. For the reasons noted above, distributions of point estimates need not be comparable when test forms are changed, even when the IRT model holds. Ad hoc methods of reconciling point estimates from different scores lead to discontinuities in trends and spurious changes in distribution features.

These problems can be circumvented by a more careful explication of the evidentiary relationship between data from one form and point estimates on another. In particular, two-stage projection expresses the evidentiary relationship between data collected from unconstrained test forms and IRT point estimates for a reference test form. The assessment designer designates some particular set of items and their associated parameters as a reference form, Test X. The designer then projects information obtained from the various data-gathering forms into the observed-score scale of the Test X. The relationships are shown in Table 8.

To illustrate the meaning of the table entries, consider the first row in the table, for the population mean on the scale of observed IRT estimates on the marketbasket form Test X, using data collected with unconstrained Test U. First, $p(\theta|u_i)$ expresses

Table 8

Relationships Between Data Collected With an Unconstrained Form and Features of the Distribution of IRT Point Estimates for a Reference Test X: Two-Stage Item-Level Model-Based Projection

Target of Inference	Two-Stage Model-Based Projection: Item-Level Model (e.g., IRT, MIRT)
Population Mean	$\overline{\overline{\hat{\theta}_X}} = N^{-1}\sum_{i=1}^{N}\int \hat{\theta}_X(x)\int p(x\vert\theta)p(\theta\vert u_i)d\theta\, dx$
Subpopulation Mean	$\overline{\overline{\hat{\theta}_{Xg}}} = N_g^{-1}\sum_{i=1}^{N_g}\int \hat{\theta}_X(x)\int p(x\vert\theta)p(\theta\vert u_i, g)d\theta\, dx$
Population Variance	$\overline{\overline{\sigma_{\theta}^2}} = N^{-1}\sum_{i=1}^{N}\int\left(\hat{\theta}_X(x)-\overline{\overline{\hat{\theta}_X}}\right)^2\int p(x\vert\theta)p(\theta\vert u_i)d\theta\, dx$
Subpopulation Variance	$\overline{\overline{\sigma_{\hat{\theta}g}^2}} = N_g^{-1}\sum_{i=1}^{N_g}\int\left(\hat{\theta}_X(x)-\overline{\overline{\hat{\theta}_{Xg}}}\right)^2\int p(x\vert\theta)p(\theta\vert u_i, g)d\theta\, dx$
Population $\mathrm{PAC}(X_0)$	$\overline{\overline{\mathrm{PAC}(\hat{\theta}_{X0})}} = N^{-1}\sum_{i=1}^{N}\int c(\hat{\theta}_X(x);\hat{\theta}_{X0})\int p(x\vert\theta)p(\theta\vert u_i)d\theta\, dx$
Subpopulation $\mathrm{PAC}(X_0)$	$\overline{\overline{\mathrm{PAC}_g(\hat{\theta}_{X0})}} = N_g^{-1}\sum_{i=1}^{N_g}\int c(\hat{\theta}_X(x);\hat{\theta}_{X0})\int p(x\vert\theta)p(\theta\vert u_i, g)d\theta\, dx$

Note. $\hat{\theta}_X(x)$ denotes the IRT point estimate calculated for response vector *x* on Test X.

what is learned from Student *i*'s responses to the items in Test U, about her value on the (possibly vector-valued) latent variable θ that underlies responses to the items of both Test U and Test X. (Obtaining this distribution once again requires Bayes theorem, and requires that $p(\theta)$ be known or previously estimated.) Next, one uses that projection into the θ metric as a weighting function to obtain the expected probability for Subject *i* to produce a given response vector *x* on marketbasket Test X; that is, $p\left(x\middle|u_i\right) = \int p(x|\theta)p\left(\theta\middle|u_i\right)d\theta$. Now $\hat{\theta}_X(x)$ is the IRT estimated score corresponding to this vector—a maximum likelihood estimate, for example. Then $\int \hat{\theta}_X(x) \int p(x|\theta)p\left(\theta\middle|u_i\right)d\theta\, dx$ is the expected value of this marketbasket IRT score, averaged over all possible response vectors, each weighted by the projected probability that Student *i* would produce that response vector in light of the Test U vector that actually was observed. Finally, $\overline{\overline{\hat{\theta}_X}}$ is the mean of such values over all sampled students.

Using such reasoning, Table 8 provides for estimates of features of IRT point-estimate distributions on a target marketbasket form Test X, given responses on an unconstrained form under the assumption that a common latent trait model governs responses on all items concerned. The remaining paragraphs in this section offer further comments that pertain to particular cells in this row of the matrix.

4.5.1 Data Collected Through Single Form

If a single form Test X is administered to all students, and IRT point estimates define the reporting scale, then no projections are required. Intuitive test theory takes over, once the IRT point estimates have been obtained student by student. While less complicated than the projection method described above, this approach falters when Test X is either very short, very hard, or very easy. In these cases the distribution of point estimates may not be sensitive to differences among subpopulations or over time points that are of interest, or capable of tracking small changes in PACs.

Nevertheless, gathering data with a single form and reporting on distributions of IRT point estimates with respect to this form present no logical flaws. It is only when the test form(s) used to gather data change that more the subtle problems arise, and more complex analytical procedures such as those of Table 8 are required to rectify the problem.

4.5.2 Data Collected Through Parallel Forms

The form-to-form vagaries of distributions of IRT point estimates are substantially lessened if the forms have very similar test information curves—a condition that can be approximated if care has been taken to construct forms that are essentially parallel with respect to content and composition. However, the more test forms depart from this condition, the more dissimilar the distributions of point estimates they will engender for a given true θ distribution. The variation of estimates of distributional features from data gathered with different forms, above and beyond variation that can be attributed to student sampling, is a gauge of the uncertainty introduced by the lack of parallelism.

4.5.3 Data Collected Through Tau-Equivalent Forms

With increasingly long tests with no floor or ceiling problems, IRT point estimates for a given student from the various forms converge to the same value; that is, IRT point estimates are consistent for any given student as test length increases. For population and subpopulation means, averages of IRT point estimates will thus tend to the same values. (Again, this is not generally true for other features of the distributions of interest.) This is a large-sample argument in terms of the length of administered test forms, however—a situation avoided for practical reasons in most large-scale educational surveys. If test forms are constructed to be tau-equivalent, acceptable approximations to this desirable result for means may occur earlier than for unconstrained forms; variations in point estimate distributions due to differences in test difficulties have been eliminated. But since the θs are estimated with different precision, other features of the distribution will not agree from form to form in any case short of perfect measurement of individual students.

4.5.4 Data Collected Through Congeneric Forms

Equal-precision adaptive tests (Wainer et al., 2000, chap. 5), which are a special kind of congeneric forms, may provide acceptable results for reporting in the metric of IRT point estimates. These are adaptive tests in which testing continues for each subject until a targeted estimate of the standard error or posterior standard deviation has been attained; thus all students have scores of about the same measurement error. What is required, however, is equally precise measurement for all students across all subpopulations that are to be compared, including comparisons over time. This condition may be hard to satisfy if the item pool contains widely diverse kinds of test items, or requires a multivariate IRT model.

Whether the approximation will be sufficiently accurate to measure small changes in a survey such as NAEP remains to be verified empirically, through simulations and field tests.

The generally applicable two-stage projections noted above as a better way to think about reporting on an IRT point-estimate scale can be used with CAT data representing Test U. Though more complex computationally, these projections would be more robust to inevitable variations in test length and accuracy, over time and across forms.

4.5.5 Data Collected Through Unconstrained Forms

As discussed at the beginning of this section, two-stage projection to observed-responses followed by point estimates for Test V is the appropriate way to reason from responses to arbitrarily constructed forms and IRT point estimates for this marketbasket form (Table 8). The more the constitution of forms varies, the more likely it is that a multivariate IRT model will be needed.

It may be noted that the IRT model for the point-estimate reporting scale need not be the same one as is used to actually model responses. For example, the five-dimensional IRT model used in the 1986 NAEP mathematics survey can be used to project responses from any data collection form to a response vectors on a reference form, but then a different, perhaps one-dimensional, IRT model can be used to summarize scores for this reference form only. This distinction is subtle, so let us be explicit about the two latent variable models that are involved here: One and the same latent variable model is assumed to *produce* item responses and to *project* information from Test U responses to the Test X response space; this is the model used for $p(\theta|u_i)$ and $p(x|\theta)$. A second, different latent variable—not necessarily believed to be responsible for generating responses—can be used to reduce projected x score vectors to point estimates; this is the model used for $\hat{\theta}_X(x)$.

4.6 Inference in Terms of a Latent Variable on the Latent Variable Scale

Reporting with respect to a latent variable θ means synthesizing information from item responses in the form of an estimated distribution onto the θ scale. Again θ can be unidimensional or multidimensional; and if it is multidimensional, a function $S(\theta)$ that produces a unidimensional summary can be defined. In all cases, though, this reporting option concerns the distribution of θ in the population of interest, as opposed to the distribution of $\hat{\theta}$. Reporting with respect to the θ scale is the reporting approach currently used in NAEP (Section 4.6.5).

With all methods of collecting data, one-stage projection to the θ scale expresses the evidentiary relationship between the item responses that are observed and the distribution on the reporting scale. Table 9 gives the equations for these relationships.

Since IRT point estimates of individual students' θs are consistent as test length increases, it is an empirical question of when sufficiently long test forms yield $\hat{\theta}$s precise enough to be treated as θs, for the purpose of estimating population features. An experiment carried out in the context of NAEP by Johnson, Liang, Norris, and Nicewander (1996) indicated that double-length NAEP forms, which required about two class periods of testing time from each sampled student, were not sufficiently accurate to achieve this result.

Further comments on particular data collection methods follow.

Table 9

Relationships Between Data Collected With Test U and Features of the Latent Variable Distribution: One-Stage Model-Based Projection

Target of Inference	One-Stage Model-Based Projection
Population Mean	$\overline{\overline{\theta}} = N^{-1}\sum_{i=1}^{N}\int \theta\, p(\theta \mid x_i)\, d\theta$
Subpopulation Mean	$\overline{\overline{\theta_g}} = N_g^{-1}\sum_{i=1}^{N_g}\int \theta p(\theta \mid x_i, g)\, d\theta$
Population Variance	$\overline{\overline{\sigma_\theta^2}} = N^{-1}\sum_{i=1}^{N}\int \left(\theta - \overline{\overline{\theta}}\right)^2 p(\theta \mid x_i)\, d\theta$
Subpopulation Variance	$\overline{\overline{\sigma_{\theta g}^2}} = N_g^{-1}\sum_{i=1}^{N_g}\int \left(\theta - \overline{\overline{\theta_g}}\right)^2 p(\theta \mid x_i, g)\, d\theta$
Population PAC(θ_0)	$\overline{\overline{\mathrm{PAC}(\theta_0)}} = N^{-1}\sum_{i=1}^{N}\int c(\theta;\theta_0)p(\theta \mid u_i)\, d\theta$
Subpopulation PAC(θ_0)	$\overline{\overline{\mathrm{PAC}_g(\theta_0)}} = N_g^{-1}\sum_{i=1}^{N_g}\int c(\theta;\theta_0)p(\theta \mid u_i, g)\, d\theta$

4.6.1 Data Collected Through Single Form

IRT-based methods for estimating the distributions of latent variables apply here in their simplest forms. If there are no background variables, for example, we can use the methods in Mislevy (1984). Incorporating increasingly many background variables G requires either more complex models for the latent distribution, incorporating perhaps many parameters for conditional distributions of θ given g as has been done in NAEP (Allen, Johnson, Mislevy, & Thomas, 1999). Alternatively, distributions can be estimates for θ conditional on just a few background variables at a time, as attention focuses on each in turn (Cohen 1997; Mislevy, 1985).

4.6.2 Data Collected Through Parallel Forms

The same issues arise here as in the previous cell. One-stage projection through an IRT model is the appropriate way to understand the underlying evidentiary relationship.

4.6.3 Data Collected Through Tau-Equivalent Forms

Again, one-stage projection through an IRT model is the appropriate way to understand this relationship.

4.6.4 Data Collected Through Congeneric Forms

One-stage projection through an IRT model is the appropriate way to understand this relationship, too. The required estimation procedures are rendered more stable and efficient in the special case of the congeneric test forms that adaptive testing represents.

4.6.5 Data Collected Through Unconstrained Forms

Once again, one-stage projection through an IRT model is the appropriate way to understand this relationship. With arbitrarily constructed forms, it is more likely that a multivariate IRT model will be needed. This is the combination of data collection and reporting scale in NAEP as this is written (Allen et al., 1996). NAEP uses a multiple-imputation approach as described in Section 4.1.5 to analyze the resulting data. Steps 6 and 7 are skipped, though, because (after a linear transformation) the draws from examinees' posterior distributions in Step 5 are immediately on NAEP's IRT true-score reporting scales.

5.0 Discussion

The preceding sections have developed the machinery necessary for thinking through the relationships among reporting scales, data collection methods, and requisite analytic procedures in large-scale educational assessments. This section offers some observations on tradeoffs they entail for assessment design—advantages and disadvantages, and strategies for attaining some competing goals.

5.1 The Inevitability of Tradeoffs

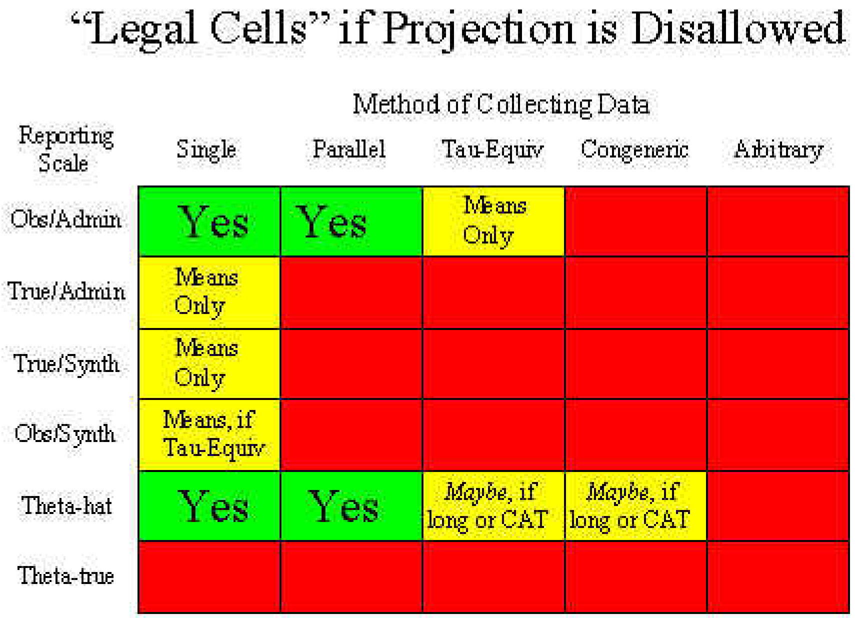
Each reporting metric represented in the rows of the matrix has advantages and disadvantages, each data-gathering method represented in the columns has advantages and disadvantages, and the analytic methods required in each cell in the intersections have pluses and minuses. Tradeoffs are inevitable because there is no single combination that dominates all others.

Marketbasket reporting metrics, for example, present two hard choices. The first is a tradeoff between comprehensibility and generalizability. The first two rows are based on a marketbasket collection short enough to administer to a student—or to publish in the newspaper, or distribute to parents. Easy to comprehend, as it fits right in to everyday experience with tests. And not very generalizable. It isn't possible for any one administerable form to cover the richness and variety of a content domain, so performance on that particular collection of tasks would overstate the performance of some groups at the expense of others, and might tend to narrow instruction if performance on this single set of tasks were taken to represent all that is important for students to learn. The third and fourth rows are based on a large and representative set of tasks. Domain coverage, in breadth and depth, can be better achieved. But it is harder to know exactly what scores on such a test represent, and one cannot test it out on oneself, one's students, or one's children.

The second tradeoff in reporting metrics contrasts the first and third rows against the second and fourth: observed-score reporting metrics versus true-score reporting metrics. Observed-score metrics are easier to understand; they fit right in to intuitive test theory. But unless data are gathered on the marketbasket form or one parallel to it (cells [1,1] and [1,2]), complex statistical machinery (two-stage projection) is required to bring information to an easy-to-interpret metric.

Data-gathering methods pose tradeoffs as well. A single form makes for the easiest test construction, administration, and potential linkage to a marketbasket metric. But a single test is easily compromised, and offers the poorest domain coverage and generalizability. Parallel tests increase coverage, but at the cost of considerable discipline in test construction. Tau-equivalent and congeneric tests offer additional flexibility for test developers and administrators, but require more complex analyses. Unconstrained forms are best from the point of view of test developers, and most esoteric from the point of view of analysts.

Reporting metrics and data-gathering methods intersect in cells to bring about evidentiary relationships between observations and desired reports. All other things being equal, simple relationships are more desirable than complex relationships, because complex relationships mean more complex analyses are required. But to simultaneously increase flexibility for test developers, comprehensibility for users, and accuracy for assessment owners will force the analyst into cells with complex evidentiary relationships. They are obscure, sometimes mistrusted, from the point of view of assessment users and even assessment owners. More assumptions are required, and there are more places in the analysis for things to go wrong, and they are harder to track down when they do. For example, if projections are eschewed because they are too esoteric, only four cells in the matrix can provide solid estimates for all population and subpopulation features (Figure 7).



“Legal Cells” if Projection is Disallowed

	Method of Collecting Data				
Reporting Scale	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin	Yes	Yes	Means Only	🟥	🟥
True/Admin	Means Only	🟥	🟥	🟥	🟥
True/Synth	Means Only	🟥	🟥	🟥	🟥
Obs/Synth	Means, if Tau-Equiv	🟥	🟥	🟥	🟥
Theta-hat	Yes	Yes	*Maybe*, if long or CAT	*Maybe*, if long or CAT	🟥
Theta-true	🟥	🟥	🟥	🟥	🟥

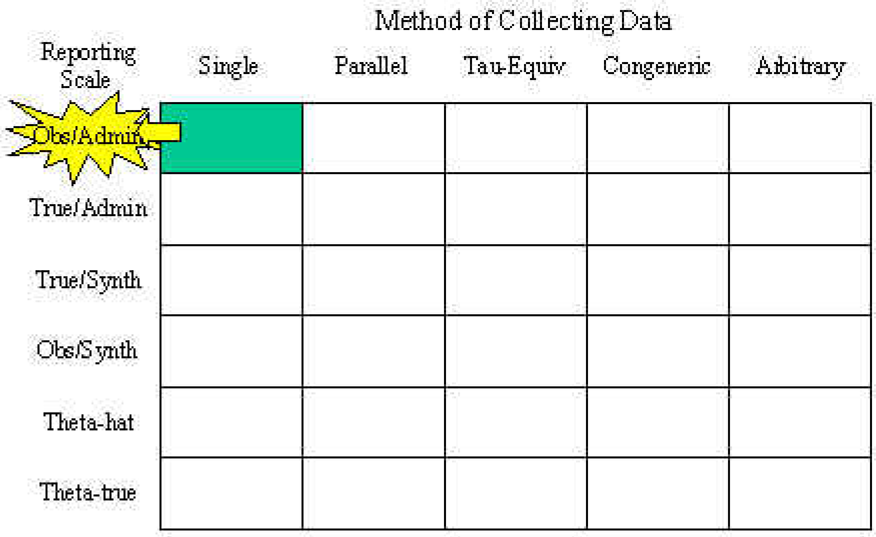
Figure 7. Summary of which cells provide correct relationships for population and subpopulation distributions if projection methods may not be used.

Suppose an assessment program gathers data and reports results in a combination that leads to one of the proscribed cells, and carries out analyses that do not use projection. They will obtain biased estimates of some or all of the features of population and subpopulation distributions in the intended reporting metric. Biased results may be tolerable in a static system, however. If data are gathered in each subpopulation and across all time points in the same way, analyzed with the same models, and reported on the same scales, then comparisons among groups and trends over time can provide useful information; intuitively, through experience, users adjust to the biases in their interpretations. Watching the percentage of students at "Advanced" move from 5% to 6% to 9% over three years differs little in any practical sense from 2% to 3% to 5%, especially in light of the uncertainty and arbitrariness of setting the Advanced cut point anyway. Problems are more serious when form composition or length changes, however, since using the same reporting scale and applying the same mismatched analyses produces incomparable results. Having the percentage of Advanced students plummet from 9% to 2% simply because more reliable test forms were administered is a disaster of a magnitude that can sink an assessment program—not to mention sapping resources from the educational system to assign blame and install remedies for a phantom problem.

5.2 Comments on Some Particular Cells

The assessment configuration that sits best with people's intuitive thinking about tests is Cell[1,1] (Figure 8): Everyone takes the same test, analysis is simply adding up number right or percent-correct for every student individually, and results are reported in terms of these student-by-student total scores. This is the simplest cell in the matrix to administer, analyze, and report—and it is also the most seriously constrained.

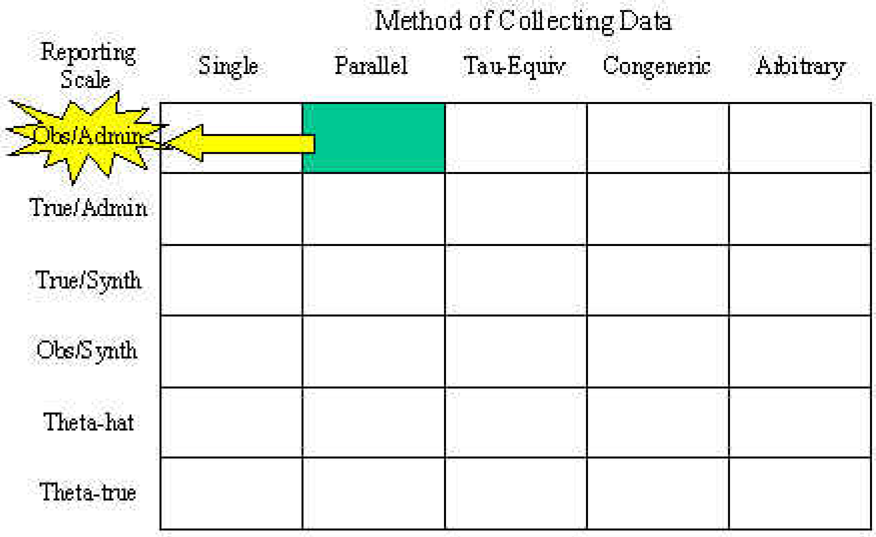
Because of experience with testing programs such as the SAT, certification exams, the Armed Services Vocational Aptitude Battery (ASVAB), and standardized achievement tests such as the Iowa Tests of Basic Skills, the public has also grown accustomed to the use of parallel test forms. Sometimes the results are reported in terms of total scores (Cell[1,2]), possibly after equating, and other times in terms of scaled scores (Cell[5,2]). Serious discipline is required on the part of test developers to produce interchangeable test forms, but the benefit of easy interpretation and straightforward analysis is maintained while increasing domain coverage and test



Reporting Scale	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin	🟩				
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

Figure 8. The configuration on which intuitive test theory is grounded—Also the easier cell in the matrix for design, analysis, and reporting.

security. Cell[1,2] (Figure 9) therefore offers some appeal for an ongoing assessment system. This might be called a short form/marketbasket reporting configuration.



Reporting Scale	Single	Parallel	Tau-Equiv	Congeneric	Arbitrary
Obs/Admin		🟩			
True/Admin					
True/Synth					
Obs/Synth					
Theta-hat					
Theta-true					

Figure 9. A configuration to which the public has grown accustomed.

A serious disadvantages of the short form/marketbasket reporting configuration is the inflexibility of form designs, both within assessment administrations and over time. Because the test forms must remain parallel and fixed in structure, the assessment cannot adapt easily to changes in the content domain over time as to kinds or balances of tasks. Further, an implicit restriction is imposed on the range of tasks that can be offered. Assessors would not have the flexibility they currently have in NAEP, for example, to use some test forms that call for some students to write long essays while others ask students to write a series of shorter passages, or in science, to present some students extended science investigations and others a collection of short-answer questions that span a range of concepts. To accommodate this flexibility, NAEP must use a multivariate IRT model and use analytic procedures appropriate for unconstrained form compositions. This places NAEP in the fifth column of our matrix. The reporting scale is that of the latent variables, locating the current NAEP in Cell[6,5] (Figure 10). This is the second most complex cell in the entire matrix, from the point of view of the analyst.

What then is the most complex cell, from the point of view of the analyst? It is Cell[1,5] (Figure 11). First, data are gathered with unconstrained forms, so that the widest possible range of tasks can be presented, changes in form length and composition can be effected within and across time, and balances of tasks can evolve with the content area. Second, results are reported with respect to observed scores on a single administerable marketbasket form, so that interpreting results is as intuitive and familiar to users as it can be. The radical difference between how data are gathered and how they are reported requires complex analyses, in the form of multivariate latent variable modeling and two-stage projection all the way to imputed responses on the marketbasket form, for students who were administered tests that looked nothing like it.

Despite its appeal to assessment users and designers, this approach harbors another deficiency: Not all of the information gathered on the unconstrained forms can be revealed on the marketbasket reporting scale. Recall that the marketbasket scale says how students, however assessed, would likely have done on the marketbasket items if they had taken them. How well they did, or might do, on in-depth inquiry investigations or extended writing samples is relevant only insofar as those tasks overlap with the marketbasket tasks. If performance is increasing with respect to inquiry skills relative to science content knowledge, but the marketbasket

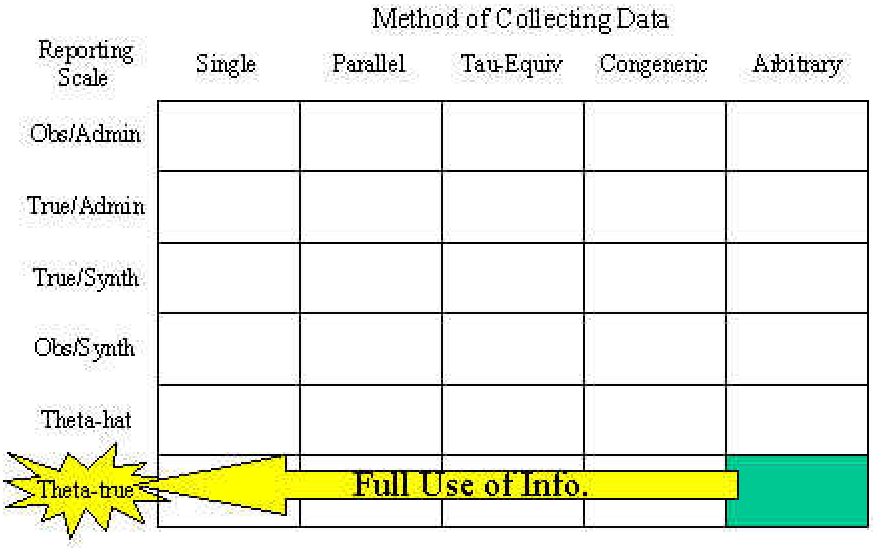


Figure 10. The current NAEP configuration.

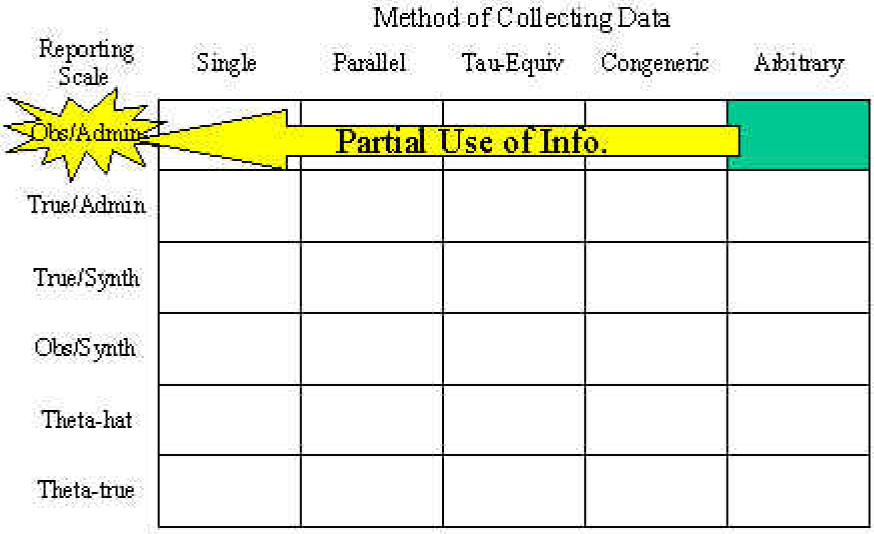


Figure 11. The most complex cell in the matrix.

contains only content knowledge items, then the unique gains on the inquiry skills cannot show up in the marketbasket reports. The next section considers ways to remedy this loss—not surprisingly, at the cost of adding complexity of one kind or another.

5.3 Using Multiple Data Collection Methods and Multiple Reporting Scales

As we have seen repeatedly, each approach to collecting assessment data and reporting results has its own advantages and disadvantages, and no single approach simultaneously best satisfies the preferences of all stakeholders, including users, designers, owners, and analysts. One strategy for designing an assessment system, then, is to employ *multiple* methods of gathering, analyzing, and reporting data in the same assessment. The idea is to employ a combination of data-gathering methods, and perhaps reporting scales as well, in a way that provides to some degree the advantages of each while compensating for its disadvantages by using an alternative approach in the same assessment.

A configuration illustrated in Figure 12 that combines two approaches discussed in Section 5.3 is halfway there. Data are gathered here in two ways. One way is with forms parallel to a marketbasket form that can be made public, so that with this approach one achieves comprehensible reports and simple analyses. The disadvantage of having to use rigid and limited parallel test forms is mitigated by the addition of unconstrained forms as well, so that a wider variety of tasks can be used and at least some test forms can vary as to content and format within and across assessments. But as noted above, not only are complex analyses needed to project results from the unconstrained forms to the marketbasket reporting scale, information that is unique to tasks not represented in the marketbasket is lost.

One way to resolve this shortcoming is to provide information about knowledge and skills outside the marketbasket in supplementary reports, as suggested in Figure 13. Advantages of such a configuration include comprehensible marketbasket reports, with some data that can be analyzed quickly and even locally in its terms, supplemented by a broader variety of data that not only contribute to the marketbasket reports but add their own information in the parts of the content domain that can’t logistically be included in the marketbasket. While the additional complexity of analyzing data from the unconstrained forms is a disadvantage, perhaps the most serious drawback is that everything that is important to learn from the assessment can’t be conveyed on a single reporting scale. Serious observers never thought that it could, of course. Yet it may seem preferable to have a primary reporting metric that captures everything and can be decomposed, rather than a primary reporting metric that effectively captures most of the information but must be supplemented by additional reports.

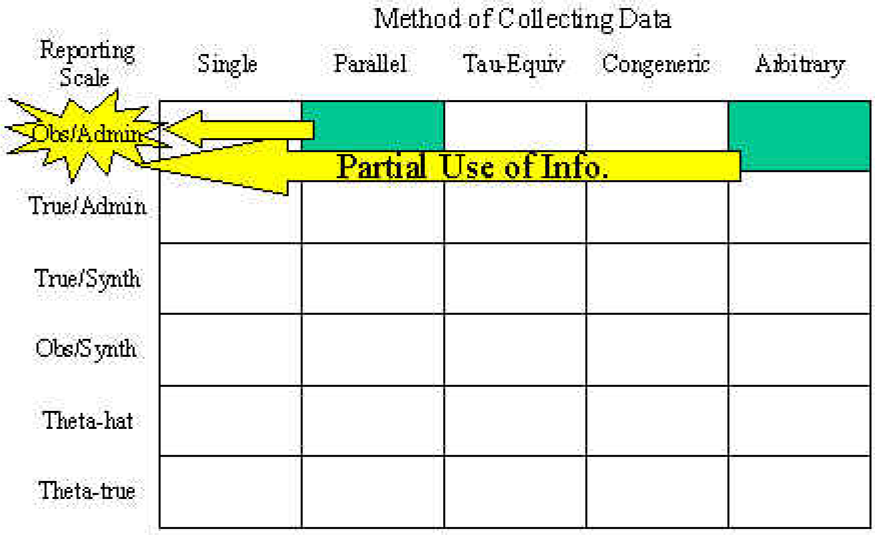


Figure 12. Two data-gathering methods providing information on the same reporting scale.

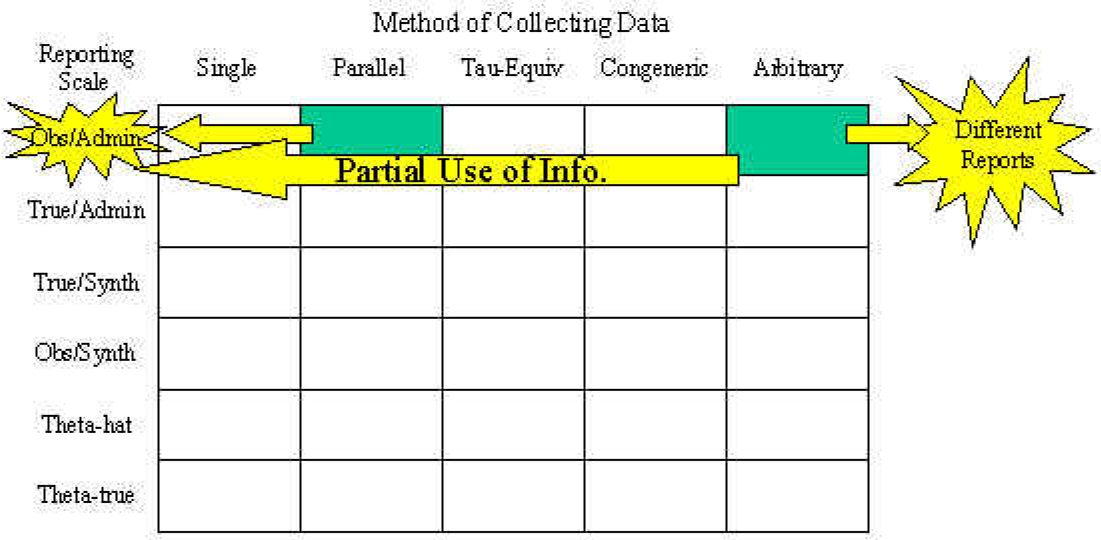


Figure 13. Observed-score marketbasket reports with supplementary reports on skills not reflected in the marketbasket.

A second way to resolve the problem is to have a complex data-gathering and reporting approach that captures all the information across the full range of items, as NAEP does now, supplemented by a streamlined marketbasket reporting scale that captures the core of that information. Figure 14 suggests such an approach, which combines (a) simple observed-score-on-an-administerable-form marketbasket reporting and parallel form data collection with (b) the current complex but comprehensive reporting of NAEP, using data from unconstrained forms to report in terms of distributions of latent variables. This option is illustrated in Figure 14. Advantages again include comprehensible marketbasket reports using data collection with locally scorable parallel forms, and a separate, more complex but fully comprehensive reporting scale based on a latent variable model. The big problem with this configuration is that certain features of distributions in the two reporting metrics may not agree with one another (Figure 15). The reasons are twofold: First, the observed-score marketbasket contains measurement error at the level of students, so, for example, a higher proportion of students will have scores above high cutoffs than the true proportion estimated on the latent variable report. Second, the latent variable report summarizes data across a broader array of knowledge and skill, so to the extent groups differ or trends vary in these skills as

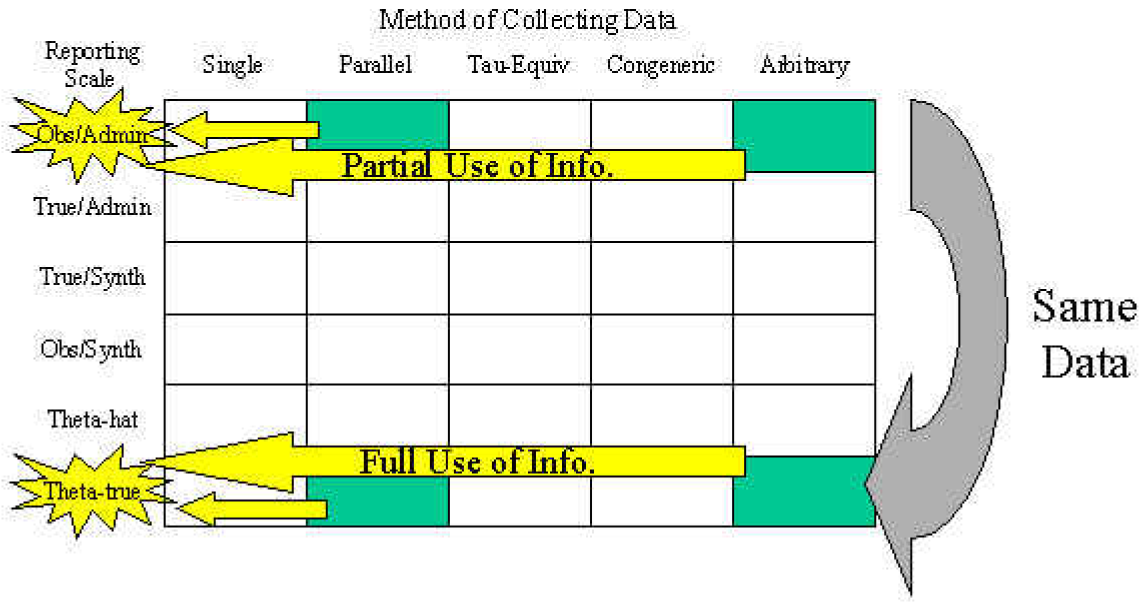


Figure 14. Dual reporting scales: Quick and comprehensible marketbasket data collection and reports, and comprehensive reports on a latent variable scale.

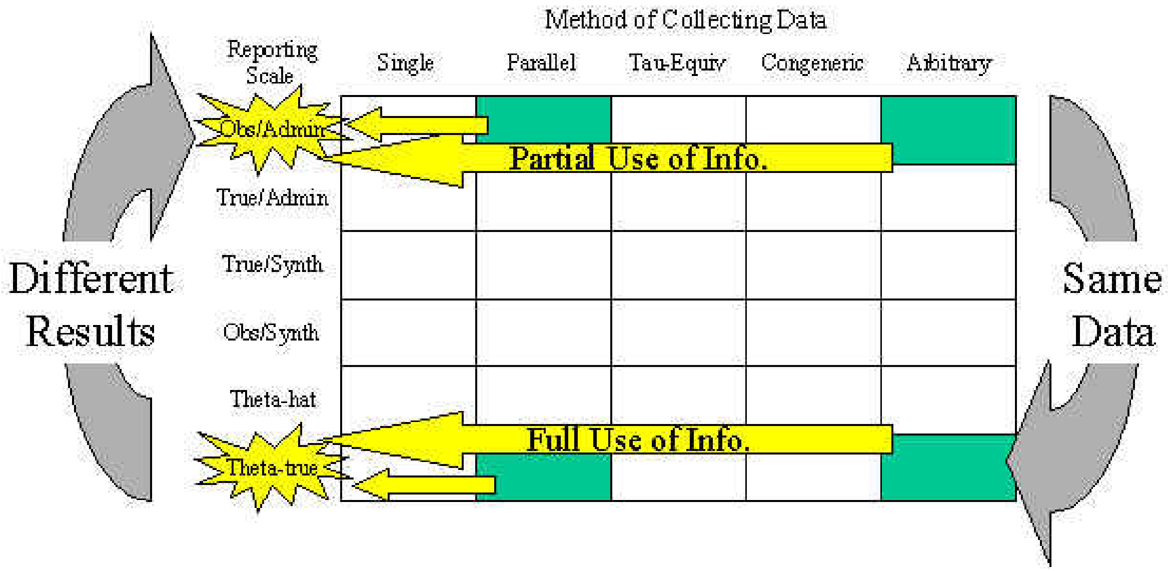


Figure 15. A potential disadvantage of dual reporting scales.

compared to the marketbasket, substantive disagreements can result. It is an open question whether users would be more happy with two numbers because they have more information and get some of it quickly and simply, or more uncomfortable because the two numbers sometimes (appropriately) tell them different things.

6.0 Conclusion

"Intuitive test theory" is based on everyone taking the same test, scores being what they are, and inferences about groups are statements about distributions of these scores (Cell [1,1]). Why do we have to bother with all these esoteric complications? Isn't IRT complicated enough already? Why can't we just use equating?

This is the reason: Once one steps beyond the simple data collection and interpretation paradigm under which intuitive test theory evolved, the noise inherent in the relationships between what one observes and what one wants to infer causes it to give the wrong answers. Paradoxes arise. Some unpleasant truths cannot be avoided. For example, except in very special cases, there is no single-point, examinee-at-a-time matchup between the evidence for a given inference about an examinee's proficiency from different test forms. The proportion of examinees with observed scores above a given level is a biased estimate of the proportion of examinees above that point. Observed-score distributions are not the same as true-score distributions, and they depend on the particulars of the test form, so different test forms give you different answers if you look at student-by-student estimates. In general, to get the right answer you have to work with entire probability distributions that capture both what you know about each examinee's proficiency and what you don't know. When targeted inferences concern populations and background variables, the probability distributions associated with a given examinee depend on her background data and the response and background data of other examinees.

The equations, the relationships, and the methodologies discussed in this paper are complex, but they are not needlessly complex. The increasingly complex techniques of test theory evolved to handle increasingly ambitious inferential problems. Each generation of developments overcomes limits on gathering and reporting assessment data that frustrate educators and policymakers. The complexity of models and methods is the price we sometimes pay for more powerful, more flexible, and more useful ways of gathering and reporting evidence about what students know and can do. For example:

- Item response theory is more complex than classical test theory, but IRT makes it possible to establish links among tests that are not parallel or tau-equivalent, to design tests that will meet targeted statistical characteristics, and to adapt tests to individual students in light of their unfolding performance.

- Matrix sampling designs are more complicated to analyze than having every student take the same test or parallel tests, but they open the door to more flexible test design and provide better coverage of educational domains.

- Projection methods are vastly more complicated than intuitive test theory, but they permit an assessment program to gather data in an efficient design yet report results on a comprehensible form.

This list could be extended much further, even if restricted to the psychometric, statistical, and survey sampling methods used in large-scale educational assessments.

This presentation has aimed to lay out the evidentiary relationships that different combinations of data collection methods and reporting methods entail. These relationships have inescapable implications for anyone who wants to report results on a particular scale, from data that arrive in a particular way: We cannot choose data-gathering methods, reporting scales, and methods of analysis independently.

Suppose we choose a data-gathering method and a reporting scale. We must then use analyses that are consistent with the evidentiary relationships that our choices entail. For example, if we decide to (a) report results on the observed-score metric of a synthetic marketbasket and (b) collect data using unconstrained test forms in order to minimize constraints on test developers, we have to use some analytic procedure that is model-based and involves projection.

Suppose on the other hand we decide that we will not use any analyses that involve projection. This eliminates every cell in the matrix where projection is the only way to understand evidentiary relationships (Figure 7), and restricts our possible targets of inference in cells where simpler methods support inferences about means but not inferences about PACs. For example, we can gather data using parallel forms and get sound reports on an observed-score marketbasket scale for one of the forms using only standard equating. But without projection, we can't draw inferences on a true-score marketbasket scale.

Understanding the evidentiary relationships between the inferences that are desired and the data that will be collected isn't just a technical problem for the technical people. The owners of an assessment are responsible for appreciating which design decisions are negotiable and which are not. Methods of data collection are open to negotiation. Reporting scales are open to negotiation. Targets of inference are open to negotiation. Given a combination of data collection method and reporting scales, the methods of analysis and computing approximations, *among those that account correctly for the evidentiary relationships these choices create,* are open to negotiation. The evidentiary relationships themselves are not.

References

Allen, N. L., Johnson, E. G., Mislevy, R. J., & Thomas, N. (1999). Scaling procedures. In N. L. Allen, J. E Carlson, & C. A. Zelenak (Eds.), *The NAEP 1996 technical report* (NCES 1999-452, pp. 235-253). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika*, *42*, 357- 374.

Beaton, A. E. (1987). *The NAEP 1983/84 technical report* (NAEP Report 15-TR-20). Princeton, NJ: Educational Testing Service.

Beaton, A. E., & Johnson, E. J. (1990). The average response method of scaling. *Journal of Educational Statistics, 15*, 9-38.

Bock, R. D. (1993). *Domain referenced reporting in large scale educational assessments.* Paper commissioned by the National Academy of Education for the Capstone Report of the NAE Technical Review Panel on State/NAEP Assessment.

Bock, R. D., Thissen, D., & Zimowski, M. E. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, *34,* 197-211.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Carlson, D. C. (1979). Statewide assessment in California. *Studies in Educational Evaluation*, *5*, 55-75.

Cohen, J. (1997). *An overview of AIR's direct estimation software for marginal maximum likelihood models.* Washington, DC: American Institutes for Research.

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

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design feasibility team report to the National Assessment Governing Board.* Washington, DC: National Assessment Governing Board.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

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

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: American Council on Education/Macmillan.

Johnson, E. G. (1996, August). *A demonstration of market-basket reporting.* Presentation to the National Assessment Governing Board, Washington, DC.

Johnson, E. G., Liang, J.-L., Norris, N., & Nicewander, A. (1996, April). *Directly estimated NAEP scale scores from double-length assessment booklets—A replacement for plausible values?* Presentation at the annual meeting of the National Council on Measurement in Education, New York.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109-133.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: World Book.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer.

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

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*, 239-270.

Lord, F. M. (1969). Estimating true score distributions in psychological testing (An empirical Bayes problem). *Psychometrika*, *34*, 259-299.

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

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

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993-997.

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.* Princeton, NJ: Educational Testing Service, Policy Information Center. (ERIC Document Reproduction Service No. ED353302)

Mislevy, R. J. (1998). Implications of market-basket reporting for achievement level setting. *Applied Measurement in Education, 11,* 49-63.

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.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics.* Reading, MA: Addison-Wesley.

National Research Council. (2000). *Designing a market basket for NAEP: Summary of a workshop.* Committee on NAEP Reporting Practices, Board on Testing and Assessment, P. J. DeVito & J. A. Koenig (Eds.). Washington, DC: National Academy Press.

National Research Council. (2001). *NAEP reporting practices: Investigating district-level and market-basket reporting*. Committee on NAEP Reporting Practices, Board on Testing and Assessment, P. J. DeVito & J. A. Koenig (Eds.). Washington, DC: National Academy Press.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.

Reckase, M. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer-Verlag.

Rosenbaum, P. R., & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology, 40,* 43-49.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Sanathanan, L., & Blumenthal, N. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association*, *73*, 794-798.

Sirotnik, K., & Wellington, R. (1974). Incidence sampling: An integrated theory for "matrix sampling." *Journal of Educational Measurement*, *14*, 343-399.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*, 277-296.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics 2*, 309-322.

Thomas, N. (2000). Assessing model sensitivity of the imputation methods of the NAEP. *Journal of Educational and Behavioral Statistics 25,* 351-372.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Erlbaum.