HOW TO EQUATE TESTS WITH LITTLE OR NO DATA

Robert J. Mislevy
Kathleen M. Sheehan
Marilyn Wingersky

This research was sponsored in part by the Cognitive Science Program
Cognitive and Neural Sciences Division
Office of Naval Research, under
Contract No. N00014-88-K-0304
R&T 4421552

Robert J. Mislevy, Principal Investigator

Educational Testing Service
Princeton, New Jersey

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

Approved for public release; distribution unlimited.
Standard procedures for equating tests, including those based on item response theory (IRT), require item responses from large numbers of examinees. Such data may not be forthcoming for reasons theoretical, political, or practical. Information about items' operating characteristics may be available from other sources, however, such as content and format specifications, expert opinion, or psychological theories about the skills and strategies required to solve them. This paper shows how, in the IRT framework, collateral information about items can be exploited to augment or even replace examinee responses when linking or equating new tests to established scales. The procedures are illustrated with data from the Pre-Professional Skills Test (PPST).
How to Equate Tests with Little or No Data

Robert J. Mislevy, Kathleen M. Sheehan, and Marilyn Wingersky
Educational Testing Service

February, 1992

This work was supported in part by Contract No. N00014-88-K-0304, R&T 4421552, from the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research. We are grateful to Neil Dorans, Charlie Lewis, Martha Stocking, and Michael Zieky, the editor, and an anonymous referee for helpful discussions and comments. The analyses of the Pre-Professional Skills Test (PPST) were supported by ETS development funds, and were carried out in collaboration with Louann Benton, Kalle Gerritz, Robin Huffman, Nancy Petersen, Clyde Reese, Duanli Yan, and June Zack.
How to Equate Tests with Little or No Data

Abstract

Standard procedures for equating tests, including those based on item response theory (IRT), require item responses from large numbers of examinees. Such data may not be forthcoming for reasons theoretical, political, or practical. Information about items’ operating characteristics may be available from other sources, however, such as content and format specifications, expert opinion, or psychological theories about the skills and strategies required to solve them. This paper shows how, in the IRT framework, collateral information about items can be exploited to augment or even replace examinee responses when linking or equating new tests to established scales. The procedures are illustrated with data from the Pre-Professional Skills Test (PPST).

Key words: Bayesian estimation, cognitive processes, collateral information, equating, item response theory
How to Equate Tests with Little or No Data

Selection and placement testing programs update their tests periodically, as the specific content of the items becomes obsolete or familiar to prospective examinees. Because the new test forms may differ in difficulty or accuracy even if they tap the same underlying skills as the old forms, some kind of “equating” or “linking” is required to compare results across forms (Angoff, 1984). Standard procedures, including those based on item response theory (IRT), require examinee responses to both new items and items already linked to an established scale. One can determine levels of comparable performance on new and old test forms to any desired degree of accuracy by increasing the number of examinees in the linking sample.

Two disparate developments in educational measurement can prevent gathering the data that standard equating procedures require. First, current legislative activity in New York is intended to limit the administration of nonoperational items in that state, including those used in pretesting and equating. Second, the growing interest in modeling the cognitive processes of solving test items (Embretson, 1985) and the capability of microcomputers to construct tasks around cognitively salient features (Bejar, 1985; Irvine, Dann, & Anderson, in press) raise the possibility of custom-building test items for each examinee on the spot.

Although operational equating procedures rely solely upon examinee responses, researchers have been aware for some time of alternative sources of information about the operating characteristics of test items. Lorge and Kruglov (1952, 1953), for example, investigated the degree to which expert and novice judges could predict the difficulties of arithmetic test items, and Guttman (1959) predicted partial orderings and relationships

---

1 If Test A is administered to Group A and Test B to Group B, the tests can be equated if either (1) tests A and B contain common items, (2) Groups A and B overlap, or (3) Groups A and B are representative samples from the same population of examinees (Lord, 1982).
among inter-item correlations between racial-attitude items constructed according to a facet
design. More recent studies with a psychometric orientation have examined the degree to
which IRT parameters can be predicted from educationally-relevant features of items (e.g.,
Fischer, 1973; Tatsuoka, 1987), and others with a psychological perspective have focused
on task attributes that are important in cognitive processing models (e.g., Whitely, 1976).
The moderate to high relationships between item features and operating characteristics are
of considerable theoretical importance, as a framework for assessing test validity and for
constructing tests around principles of learning and knowing.

But moderate to high relationships between item features and operating
characteristics are the information equivalent of small to moderate examinee sam-
(Mislevy, 1988)—too little for standard large-sample equating procedures to work
properly. And when it comes to test equating, collateral information differs from response-
data information in a crucial respect: Linking information from examinee responses can be
made arbitrarily accurate by increasing the sample size, but information from collateral data
is limited by the strength of its relationship to item operating characteristics. Procedures
have not been available to provide coherent inferences about item operating characteristics,
and the equating and linking functions they imply, from data that contain substantially less
information than large samples of responses.

The present paper attacks this problem for domains in which (i) an IRT model fits
reasonably well, (ii) available collateral information about test items is correlated with their
IRT parameters, and (iii) a start-up data set is available from which to build predictive
distributions for item parameters, given this collateral information. The key idea is the
treatment of the uncertainty associated with the parameters of the new items. The following
section reviews IRT test equating and linking with known item parameters. Sources of
collateral information, and ways to bring it into the IRT framework, are then discussed.
An example from the Pre-Professional Skills Test (PPST) is introduced. Linking and
Equating procedures are then extended to the case of imperfect knowledge about item parameters, and illustrated with the PPST data.

**IRT Linking and Equating**

An item response theory (IRT) model gives the probability that an examinee will make a particular response to a particular test item as a function of unobservable parameters for that examinee and that item (Hambleton, 1989). This paper addresses scalar parametric models for dichotomous test items, but the ideas apply more generally. Define $F_j(\theta)$, the item response function for Item $j$, as follows:

$$F_j(\theta) = P(X_j=1|\theta, \beta_j),$$

where $X_j$ is the response to Item $j$, 1 for right and 0 for wrong; $\theta$ is the examinee ability parameter, and $\beta_j$ is the (possibly vector-valued) parameter for Item $j$. Our example uses the 3-parameter logistic IRT model:

$$F_j(\theta) = c_j + (1-c_j) \Psi[\alpha_j(\theta-b_j)];$$

where $\Psi$ is the logistic distribution function, or $\Psi(t) = (1+\exp(-t))^{-1}$, and $\beta_j=(a_j,b_j,c_j)$ conveys the sensitivity of Item $j$, its difficulty, and the tendency of examinees with very low values of $\theta$ to answer it correctly. Under the usual IRT assumption of local or conditional independence, the probability of a vector of responses $x=(x_1,...,x_n)$ to $n$ items is the product over items of terms based on (1):

$$p(x|\theta,B) = \prod_{j=1}^{n} F_j(\theta)^{x_j}[1-F_j(\theta)]^{1-x_j},$$

where $B=(\beta_1,...,\beta_n)$.

**IRT Linking and Equating when Item Parameters are Known**

If item parameters were known, one way to compare performances on different tests would be to make inferences on the $\theta$ scale, using an estimator such as the maximum
likelihood estimate or one of the Bayesian estimates described below. The varying degrees
of difficulty and accuracy among test forms are accounted for by the different parameters of
the items that comprise them. Equation (2) is interpreted as a likelihood function for $\theta$,
$L(\theta|x,B)$, once $x$ has been observed. The value of $\theta$ that maximizes $L$ is the maximum
likelihood estimate (MLE) $\hat{\theta}$. Its variance, $\text{Var}(\hat{\theta} | \theta, B)$, can be approximated by the second
derivative of log $L$ evaluated at $\hat{\theta}$. The posterior density of $\theta$ with respect to the prior
density $p(\theta)$ is obtained as
\[
p(\theta|x,B) \propto L(\theta|x,B) p(\theta).
\]
(3)
The mean of (3) is the Bayes mean estimate $\bar{\theta}$; the variance, $\text{Var}(\theta|x,B)$, indicates the
remaining uncertainty. The mode of (3) is the Bayes modal estimate $\tilde{\theta}$.

Alternatively, the IRT model can be used to generate an equating function between
number-right or percent-correct scores on two tests, through “IRT true-score test equating”
(Dorans, 1990; Lord, 1980). The expected number-right score on Test A for an examinee
with proficiency $\theta$ is given by
\[
\tau_A(\theta) = \sum_{j \in S_A} p(x_j=1|\theta, \beta_j) = \sum_{j \in S_A} F_j(\theta),
\]
(4)
where $S_A$ is the set of indices of items that appear in Test A. The expected score on Test
B, $\tau_B(\theta)$, is defined analogously. Scores on two tests are “true-score equated” if they are
expected values of the same value of $\theta$, and the IRT true-score equating line is the plot of
all pairs of equated Test A and Test B true scores: $((\tau_A(\theta), \tau_B(\theta)))$ for $\theta \in (-\infty, +\infty)$.2

Note that the averaging that occurs in (4) is for fixed $\theta$, over the uncertainty associated with
the observational setting. Specifically, the uncertainty in scores for a given $\theta$ in standard
IRT true-score equating is the 0 or 1 for each $x_j$, with $\beta_j$ assumed known.

---

2 Under the 3PL, this relationship does not give equatings for scores below the sum of the
cjs on a given test. The practical solution is generally to extend the relationship from the
lowest point on the true-score equating curve linearly down to (0,0).
Item Parameter Estimation

But item parameters are never known with certainty; they must be estimated from observable data of one kind or another—in practice, almost always from samples of examinee responses. Bayesian inference about \( \theta \) (e.g., Mislevy, 1986; Tsutakawa & Lin, 1986) begins with a (possibly uninformative) prior distribution \( p(\theta) \), a known or concurrently estimated examinee population density \( p(\theta) \), and a response matrix \( X=(x_1,...,x_N) \) from a sample of \( N \) independently-responding examinees. The posterior distribution of \( \theta \) is

\[
p(\theta|X) \propto p(\theta) L(\theta|X),
\]

where \( L(\theta|X) \) is the marginal likelihood function for the item parameters (Bock & Aitkin, 1981):

\[
L(\theta|X) = \prod_{i=1}^{N} \int p(x_i|\theta_i,\theta) p(\theta_i) \, d\theta_i.
\]

One can obtain Bayes mean estimates \( \theta \) or Bayes modal estimates \( \hat{\theta} \), and a posterior variance matrix \( \Sigma_{\theta} \) from (5), leading to the approximations \( p(\theta|X) \sim N(\theta,\Sigma_{\theta}) \) or \( N(\hat{\theta},\Sigma_{\theta}) \). Alternatively, one obtains the MLE \( \hat{\theta} \) by maximizing (6) with respect to \( \theta \).

The consistency of \( \hat{\theta} \), \( \hat{\theta} \), and \( \hat{\theta} \) as estimators of \( \theta \) justifies using item parameter estimates from large samples of examinees as if they were known true values in IRT linking and scaling; e.g., using \( L(\theta|x,\theta=\hat{\theta}) \) for \( L(\theta|x,\theta=B) \) when estimating \( \theta \), or \( p(x_j=1|\theta,B=\hat{\theta}) \) for \( p(x_j=1|\theta,B) \) when calculating \( \tau_A(\theta) \) and \( \tau_B(\theta) \) in equating (Lord, 1982).

If \( \theta \) is not well determined—i.e., \( p(\theta|\text{data relevant to } \theta) \) is too spread out to be approximated by a single-point density—this approximation understates the uncertainty associated with subsequent inferences, and, as we shall see, can yield biased estimates.

---

3 Independent priors are typically posited for \( B \) and \( \theta \). Independent and identical priors are also posited for examinees in this presentation, but see Mislevy and Sheehan (1989a) on the role of collateral information about examinees in item parameter estimation.
"Data relevant to B" can be examinee responses (X), collateral information about the items (Y), or both. B is poorly determined when the examinee sample is small, or when only collateral information about the items is available. The preceding paragraphs addressed p(B|X); the following section addresses p(B|Y) and p(B|X,Y). We then return to methods for dealing with uncertainty about B in linking and equating.

**Collateral Information about Items**

This section discusses potential sources of collateral information (yj) about a test item, and suggests ways to express this information in terms of distributions for the item parameters βj. We assume the existence of a start-up data set in which both collateral information and item parameter estimates are available from a collection of items. The basic steps are as follows:

1. Identify features of items that are useful in predicting item operating characteristics.
2. Characterize, analytically or empirically, distributions p(β|yj) based on data from the previously administered items.
3. Employ the distributions obtained in Step 2 as prior distributions for the βs of new items, conditional on their collateral data.

**Sources of Collateral Information**

*Expert Judgment.* Irving Lorge and his students studied the degree to which experts' predictions of item difficulty could be used to construct parallel test forms (Lorge & Kruglov, 1952, 1953; Tinkelman, 1947). Raters turned out to be good at predicting the relative difficulties of items, but not absolute levels of difficulty. Thorndike (1982) found that pooled judgements from 20 trained raters accounted for between 55- and 71-percent of the variance in item difficulties in three aptitude tests—too low, he concluded with disappointment, to substitute for pretesting, say, a thousand examinees. In Chalifour and Powers' (1989) study of analytical reasoning items in the Graduate Record Examination (GRE), an experienced item writer's predictions accounted for 72-percent of
normalized item difficulty variance. Bejar (1983) found item writers' predictions accounted for only about 20-percent of the variation among difficulties and among item-test correlations in an English Usage test, and less still in a Sentence Correction test. In a subsequent study of analogy items, test developers' predictions accounted for 43-percent of the variance among item difficulties (Enright & Bejar, 1989).

Test Specifications. Educational tests are written to tap skills and knowledge in a domain of content. Osburn (1968) and Hively, Patterson, and Page (1968) suggested building "item forms," or templates to create items, around the important features of a content domain. Researchers have developed numerous taxonomies to elucidate the content domains that tests address (e.g., Mayer, 1981; Chaffin & Peirce, 1988). Test specifications can also address item formats or modalities. Because they are integral to the test development process, content and format specifications constitute a readily available source of collateral information about items. Whitely (1976) accounted for 31-percent of the variance among percents-correct of verbal analogy items with a taxonomy of types of relationships. Drum, Calfee, and Cook (1981) accounted for between 55- and 94-percent of the variance in percents-correct in 18 reading tests with "surface features" such as proportion of content words in stems, length of distractors, word frequencies, and syntactic structures. Chalifour and Powers (1989) accounted for 62-percent of percents-correct variation and 46-percent of item biserial correlation variation among GRE analytical reasoning items with seven predictors, including the number of rules presented in a puzzle and the number of rules actually required to solve it.

Cognitive Processing Requirements. From the psychologist's point of view, the salient features of an item concern the operations, strategy requirements, or working memory load of anticipated attempts to solve it. Scheuneman, Gerritz, and Embretson (1989) accounted for about 65-percent of the variance in item difficulties in the GRE Psychology Achievement Test and the Reading section of the National Teacher Examination with variables built around readability, semantic content, cognitive demand,
and knowledge demand. Mitchell (1983) derived collateral information variables from theories of cognitive processes for the Word Knowledge (WK) and Paragraph Comprehension (PC) tests of the Armed Services Vocational Aptitude Battery (ASVAB), and used them to predict Rasch item difficulty parameters. The proportions of item difficulty variance accounted for in three ASVAB forms ranged from 17- and 30-percent for WK, and from 66- to 90-percent for PC.

**Characterizing Item Parameter Distributions**

Procedures for incorporating collateral information \( y_j \) about test items in IRT include Scheiblechner (1972) and Fischer’s (1973) Linear Logistic Test Model (LLTM) and Mislevy’s (1988) extension of it. The LLTM is a 1-parameter logistic (Rasch) IRT model in which item difficulty parameters are linear functions of effects for key features of items:

\[
\beta_j = \sum_{k=1}^{K} y_{kj} \eta_k ,
\]

where \( \beta_j \) is the difficulty parameter of Item \( j \); \( \eta_k \) is the contribution of Feature \( k \) to item difficulty, for \( k=1,\ldots,K \) salient item features; and \( y_{kj} \), a known collateral information variable, signifies the extent to which Feature \( k \) is represented in Item \( j \). In Fischer’s (1973) calculus example, the collateral information about Item \( j \) was a vector of indicator variables \( y_{kj} \), for \( k=1,\ldots,7 \), denoting whether or not each of seven differentiation rules was required in its solution.

Fischer and Formann (1982) list many applications of the LLTM in which meaningful item features account for substantial proportions of item-difficulty variance, but they note that the original goal of explaining all the variation among item difficulties is never met in realistic applications. Mislevy (1988) extended the LLTM to allow for variation of difficulties among items with the same salient features, by incorporating residuals around the LLTM estimate with variance \( \phi^2 \). If the prediction model is built using
a large number of previously-calibrated test items, a predictive distribution for the difficulty parameter of a new item might thus be approximated as
\[ p(\beta_j | y_j) = \mathcal{N}\left( \sum_{k=1}^{K} y_{kj} \hat{\eta}_k, \phi^2 \right), \]
where \( y_j = (y_{1j}, \ldots, y_{Kj}) \). The mean of the predictive distribution, \( \bar{\beta}_j = \sum y_{kj} \hat{\eta}_k \), is essentially the LLTM point estimate for \( \beta_j \). Note that information about new items from collateral data can be combined with examinee responses to the same items via (5), as an informative prior distribution, to yield \( p(B|X,Y) \).

An Example from the PPST (Part 1)

The Pre-Professional Skills Test (PPST) is used to measure the reading, mathematics, and writing skills of prospective teachers during their college years. Our example concerns the reading tests from eight test forms administered between 1985 and 1990. Each form comprised forty items, although one or two items were excluded from each form due to problems with the item or the scoring key. In accordance with the item overlap design used in the PPST, nearly all of the items on the first form appeared in one or more later forms; the last two forms each had twenty unique items. A "baseline" calibration of the 144 unique items was carried out under the 3PL with a sample of approximately 5000 examinees per form, using Mislevy and Bock's (1983) BILOG program. A second "operational" calibration was carried out with a sample of only 500 examinees each for the first seven forms only, using only the 103 items that did not appear on the eighth form. This example employs a collateral information model built on the seven-form operational data to link the eighth left-out form to the operational scale. The results obtained with the baseline calibration are the standard of evaluation. Part 1 summarizes the building of the collateral information model, and demonstrates the shortcomings of using the resulting point estimates of item parameters as if they were known true values.
The conditional distributions of estimated item parameters in the seven-form operational calibration were approximated with a multivariate multiple regression model. The dependent variable was the item parameter vector (slope, intercept, lower asymptote), or $\beta_j = (a_j, -(b_j/a_j), c_j)$, with a sample size of 100 items. An initial set of 30 collateral variables consisted of codings of items' content and cognitive processing features, as proposed by a team of test developers familiar with the PPST. Two test developers rated all items from all eight forms; the averages of their ratings were employed throughout. The collateral variables included in the final prediction model were determined from separate step-down regression analyses on $a_j$, $-(b_j/a_j)$, and $c_j$. For the predictors included in the final model, descriptive summaries of the variables, proportions of rater agreement, and the parameter values in the final multivariate regression model appear in Table 1.

The proportions of variance accounted for by the prediction model were .02, .24, and .05 for the slope, intercepts, and asymptotes. This corresponds to multiple R's of .14, .49, and .22. Figure 1 plots $a$, $b$, and $c$ predictions for the 39 Form 8 items against the baseline values. Considerable variation remains for individual item difficulty ($b$) parameters, and the predictions for $a$ and $c$ parameters differ only negligibly from their averages. Figure 2 presents the test characteristic curves (TCCs) for Form 8 as constructed from the predictions and the baseline values. The TCCs give expected scores in the percent-correct metric as a function of $\theta$. Much of the noise apparent in Figure 1 has been "cancelled out" in Figure 2, as the predicted TCC is surprisingly close to the baseline TCC. The discrepancy is systematic, however. Because only 24-percent of the variance among item difficulties has been accounted for, estimates of the item difficulty point estimates are too close to their mean. Items are modeled as more similar than they really are, causing the predicted TCC to rise too sharply in this region. This problem affects the IRT true-score equating. Figure 3 shows an equating curve based on operational estimates for Form 7 and
Equating with Little or No Data
Page 11

prediction-based point estimates for Form 8, along with the curve obtained using baseline item parameter estimates for both tests.

[Insert Figures 1-3 about here]

MLEs for \( \theta \) and standard errors were calculated for a random sample of 250 examinees from Form 8, using baseline item parameters and prediction-based point estimates. Figure 4 shows the \( \hat{\theta} \). A bias corresponding to the discrepancies in the TCCs is apparent, especially at the higher end of the distribution. The scatter of the prediction-based \( \hat{\theta} \)s around their baseline counterparts reflects increased uncertainty due to incomplete information about item parameters, since the only difference between the two sets of estimates is the item parameters used to calculate them. This variance is about .10. Figure 5 shows the relative change in modelled standard errors, or square roots of the variance estimates \( \text{Var}(\hat{\theta}_i B, B) \), when calculated with prediction-based point estimates of item parameters in place of \( B \) as opposed to baseline values. The average change, about zero\(^4\), is misleading, because the actual standard error of the \( \theta \) estimates should be larger; simply calculating \( \text{Var}(\hat{\theta}_i B, \overline{B}) \) with \( \overline{B} \) in place of \( B \) neglects uncertainty about \( \theta \)s due to the remaining uncertainty about item parameters. We shall see that ignoring this uncertainty causes posterior variances for \( \theta \)s to be underestimated by about a third in this example.

Up to this point, we have seen that collateral variables do provide potentially useful information about item parameters. A test characteristic curve and \( \hat{\theta} \)s calculated with predicted item parameters, or \( \hat{\beta} \)s, are surprisingly good, given that multiple Rs for slopes, intercepts, and lower asymptotes were only .14, .49, and .10. But the shortcomings of these "best estimate" point predictions for item parameters are serious enough to prevent us from simply using them as if they were true \( \beta \) values. Biases in \( \hat{\theta} \)s appear because the \( \hat{\beta} \)s are too clustered around their average. More seriously, disregarding the uncertainty about item parameters causes substantial underestimation of the uncertainty about \( \theta \)s. In this

\(^4\) The curvature is due to the clustering of predicted item difficulties around their average.
example, a variance component of .10, about half the average of the usual error variance estimate for $\hat{\theta}s$, is being ignored.

[Insert Figures 4 & 5 about here]

**IRT Linking and Equating when Item Parameters Are Not Known with Certainty**

Consider inferences about $\theta$ with imperfect knowledge about $B$, conveyed through $p(B|x)$, where “data” refers to a calibration-sample $X$ of responses from $N$ examinees, collateral information about items, or both. The probative value about $\theta$ from $x$ is now expressed through what is sometimes called an average likelihood function, which accounts for uncertainty about $B$ by averaging over its distribution:

$$L(\theta|x, \text{data concerning } B) = \int L(\theta|x, B) p(B|x, \text{data concerning } B) dB.$$  

Tsutakawa compared Bayesian inferences about $\theta$ using $p(B|X)$ and $B=B$, under the 2- and 3-parameter logistic models (the 2PL and 3PL). Under the 2PL, the more accurate estimates of $\text{Var}(\theta|x)$ using $p(B|X)$ were higher than the usual approximation, $\text{Var}(\theta|x, B=B)$, by an average of 4 percent with $N=400$, and up to 30 percent with $N=100$ (Tsutakawa & Soltys, 1988). Under the 3PL with $N=400$, increases ranged from 50 percent to over 1000 percent in unfavorable cases (Tsutakawa & Johnson, 1990).

Similarly, uncertainty about item parameters must be taken into account in IRT true-score equating. For a fixed value of $\theta$, knowledge about the observed score distribution must take into account uncertainty about item parameters as well as uncertainty about item responses. This requires integrating over $p(B|x)$ in (4) to obtain expected scores:

$$\tau_A^{*}(\theta) = E_B[\tau_A(\theta)] = \sum_{j \in S_A} \int p(x_j=1|\theta, \beta_j) p(\beta_j|x) dB_j.$$  

The IRT true-score equating line now matches values of $\tau_A^{*}(\theta)$ and $\tau_B^{*}(\theta)$.
We note in passing that this extended definition of IRT true-score equating is consistent with a familiar practice from true-score test theory: treating total scores with the same value as equivalent when tests are random samples of items from the same pool. "True score" in this case is defined as expected percent-correct in the pool, which is naturally the expected percent-correct in a random sample of items. The fact that some samples of items will be harder than others is accounted for by adding a between-forms variance component to statements about the precision of student scores (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). This component can be reduced if, instead of simple random sampling, stratified sampling according to content specifications is used to select items; that is, prespecified numbers of items are selected from "bins" of similar items.

Items may not be literally drawn from an existing pool, but conceptually sampled through the process of writing tests to the same content specifications. This presentation extends the idea to tests constructed with possibly different numbers of items from different bins.

Numerical procedures to carry out the integration required in (7) and (8) include the second-order approximation Tsutakawa used and Rubin's (1987) multiple imputations, a variant of Monte Carlo integration (Mislevy & Yan, in press, apply this technique to uncertainty about item parameters). The current presentation employs Lewis's (1985) "expected response curve" approach, which is now described below.

**Expected Response Curves**

In dichotomous IRT models, the expected value of a correct response to Item $j$ given $\theta$ and $B$ is $F_j(\theta) = P(x_j=1|\theta, \beta_j)$. If $\beta_j$ is only partially known, through $p(\beta_j|\text{data})$, the probability of a correct response conditional on $\theta$ but marginal with respect to $B$ can be written as

$$F_j^*(\theta) = E_{\beta_j}[F_j(\theta)] = \int P(x_j=1|\theta, \beta_j) p(\beta_j|\text{data}) \, d\beta_j,$$

an "expected response curve" that gives the probability of correct response conditional on $\theta$ taking into account uncertainty about $\beta_j$ (Lewis, 1985).
Even though $F_j^*(\theta)$ is the expected value of a correct response at each value of $\theta$, it is not the same as $F_j(\theta)$ evaluated with the expected value of $\beta_j$. The shape of $F_j^*$ depends on the shape of $F_j$ and the shape of $p(\beta_j)$; in general, $F_j^*$ and $F_j$ will not be of the same functional form. A simple example in which they are may aid intuition. Suppose that $F_j$ is 2-parameter normal (2PN) with slope parameter $a_j$ and difficulty parameter $b_j$; $a_j$ is known with certainty; and $p(b_j|data)$ is $N(\bar{b}_j, \sigma^2_j)$. Then $F_j^*$ is also 2PN, but with $b_j^* = \bar{b}_j$ and 

$$a_j^* = (a_j^2 + \sigma^2_j)^{-1/2}.$$ 

In this special case, the location parameter, $b_j^*$, has the same value as the Bayes mean estimate for $b_j$. The slope parameter, $a_j^*$, is attenuated to account for uncertainty about $b_j$.

Figures 6 and 7 illustrate the situation. Figure 6 concerns a 2PN curve whose slope is known to be 1 and the whose location is known only up to $p(b) = N(0,1)$. The shaded region suggests this uncertainty with bands drawn at one and two standard deviations around the curve defined by $b=\bar{b}=0$. This central curve thus corresponds to the best estimate of $b$ under squared error loss. Also shown is $F^*$, which is also a 2PN response curve, and is also centered at 0, but with $a=\sqrt{.5}=.7071$. The attenuation toward a probability of .5 can be understood from Figure 7, a slice of the posterior distribution for $P(x=1|\theta, b)$ at $\theta=1$ as $b$ ranges from $-\infty$ to $+\infty$. As a result of uncertainty about $b$, the distribution for the probability of a correct response ranges from 0 to 1. Its mean, which is required in (8), is lower than the probability associated with the most likely value of $b$ due to the skew. The mean is shifted toward .5, landing, by definition, at $F^*(1)$.

If the information about items is independent—that is, $p(B|data) = \prod p(\beta_j|data)$—then inferences about $\theta$ that take uncertainty about $B$ into account have the same conditional independence form as when item parameters are known:

$$p(x|\theta, data concerning B) = \prod_{j=1}^{n} F_j^*(\theta)^{x_j} \left[1-F_j^*(\theta)\right]^{1-x_j} \ldots$$

(9)
After $x$ is observed, (9) can be interpreted as an expected likelihood function for $\theta$, say $L(x|\theta, \text{data concerning } B)$, or $L(x|\theta)$ for short. The posterior $p(\theta|x)$ is proportional to $L(x|\theta) p(\theta)$, and posterior means and variances for $\theta$ are obtained as usual, except they take uncertainty about $B$ into account by using $F_j^*$s rather than $F_j$s.

Equation (9) proves useful even if $p(B)$ is not independent over items. Although the dependencies among items are ignored, (9) is an example of what Arnold and Strauss (1988) call a “pseudo-likelihood;” under mild regularity conditions on the $F_j^*$s, its maximum is a consistent estimator of $\theta$. Thus for large $n$, Bayesian and likelihood point estimates of $\theta$ based on (9) have the correct expectation. Indicators of their uncertainty based on (9), however, such as the variance estimator of $\theta$ and the posterior variance, tend to be too optimistic. But if the dependencies among item parameter estimates are small—and they tend toward zero as test length increases (Mislevy & Sheehan, 1989b)—the underestimation of uncertainty about $\theta$ from this source is minor.

Expected response curves can also be used for IRT true-score equating, with

$$\tau_A^*(\theta) = \sum_j F_j^*(\theta).$$

Since only expectations are involved, (10) is correct whether or not $p(B)$ is not independent over items.

Closed-form solutions for $F^*$ are not generally available. One way to approximate $F_j^*$ is outlined below.

1. Lay out a grid of $\theta$ values across the range of interest. Denote by $\Theta_m$ the $m^{th}$ grid point.
2. For Item $j$, draw a sample of $S$ item parameter values from $p(\beta_j|\text{data})$. Denote by $\beta_j^{(s)}$ the $s^{th}$ such draw.
3. Evaluate the probability of a correct response to Item $j$ at $\Theta_m$ using each $\beta_j^{(s)}$ in turn, or $P(x_j=1|\theta=\Theta_m, \beta_j=\beta_j^{(s)})$. Denote the result $P_{jm}^{(s)}$.
4. The point on the expected response curve for $\theta=\Theta_m$ is approximated by the average of the values obtained in Step 3:
Steps 2 and 3 generate an empirical approximation of the predictive distribution of \( P(X_j=1|\theta, \beta_j) \) over the range of \( \beta_j \) for fixed values of \( \theta \), an example of which appeared as Figure 7. Step 4 is finding the posterior mean for \( P \) with respect to \( \beta_j \) conditional on each of the \( \theta \) points—approximations of the values on the expected response curve. Subsequent inferences about \( \theta \) can be drawn using these values directly in a discrete approximation of integrals involving \( \theta \) distribution, or after fitting a smooth curve to them.

It is convenient operationally to approximate each \( F^* \) with the closest curve from a familiar family—for example, the closest 3PL curve in applications based on the 3PL model, or the closest 2PL model in applications based on the 1PL or 2PL. This approach makes it possible to use standard software designed for popular parametric IRT models to estimate examinee scores, construct tests, or draw equating lines; the only difference is entering item parameters for expected response curves rather than very precise estimates of true item parameter values. Let \( F^{**} \) denote the target approximation. Given \( F^* \), a weighted least squares estimate of \( F^* \) is obtained by minimizing the fitting function

\[
\sum_{m=1}^{M} \left[ F^{**}(\theta_m|B^{**}) - F^*(\theta_m) \right]^2 W(\theta_m)
\]

with respect to the parameter \( B^{**} \) of \( F^{**} \), where \( W(\theta_m) \) is a weighting function that specifies the relative importance of matching \( F^{**} \) to \( F^* \) at various points along the \( \theta \) scale.

In practical work, one might create simulated examinees at each \( \theta_m \)-point in numbers that reflect the relative importance of fitting \( F^{**} \) at those points and with the proportion \( F^*(\theta) \) of them with correct answers in each group, then run a logit regression analysis or the LOGIST computer program (Wingersky, 1983) with the "fixed \( \theta \)" option to estimate the parameters \( B^{**} \) of a best-fitting 2PL or 3PL. Additional information that becomes available over time, say, as examinee responses are acquired in operational testing, can be incorporated merely by updating item parameter values under the same model.
An Example from the PPST (Part 2)

Expected response curves for the items of Form 8 were constructed from the predictive distributions built in Part 1 of the example, with 100 draws of \((a_j, -(b_j/a_j), c_j)\) for each item. Multivariate normal distributions were employed for each item, with means given by the multiple regression equations and the covariance matrix shown in Table 1. At each point in a \(\theta\) grid from -3 to +3 in steps of .2, the average modelled percent-correct was evaluated from each of the 100 plausible values of \(\beta_j\). The average of these values across the grid constituted a discrete, nonparametric estimate of an item’s expected response curve. For each item, the parameters of best-fitting 3PL curves were obtained using the method outlined in the preceding section.

Figure 8 shows, for eight representative items, nonparametric expected response curves and trace lines generated from baseline item parameters, point estimates from collateral information, and from parameters of 3PL fits to expected response curves. Three observations can be made from these tracelines, and similar ones for the rest of the items:

1. None of the approximations is impressive as an estimate of the baseline curve, although again it is their performance as an ensemble that counts.

2. The expected response curves are noticeably shallower than the trace lines based on point estimates. The uncertainty about the item parameters engenders this “hedging of bets.”

3. The 3PL approximations capture the nonparametric approximations quite well. From this point, we therefore refer to the 3PL fits as expected response curves.

It is essential to remember that “getting good item parameter estimates” is not our objective; rather, it is to express what we know about item parameters in a way that gives us good subsequent inferences that involve the unknown item parameter values.

[Insert Figure 8 about here]

Figure 9 shows the test characteristic curves corresponding to the baseline estimates and the expected response curves. The bias in the TCC in Figure 2, caused by the
shrinkage of the point estimates of item response curves to their means, has been largely eliminated. Similar improvements are made in reducing bias for MLEs, as can be seen by comparing Figure 10 with Figure 4. Figure 11, which should be compared with Figure 3, shows the improvement in the estimated true-score equating line between Form 8 and Form 7. Figure 12 shows the test information curves (TICs) corresponding to the baseline item parameter estimates, the point predictions generated in Part 1 of the example, and the expected response curves. The reciprocals of the values on these curves are approximate squared standard errors for MLEs of $\theta$s along the x-axis. The TIC based on point predictions, because it ignores uncertainty about item parameters, is misleadingly high—even higher than the TIC based on baseline estimates in the region where the predicted difficulties are centered. The TIC based on expected response curves is appropriately lower—about 33-percent lower than the baseline TIC on the average. Figure 13 shows the proportional increase in the standard errors of the 250 examinees. Since information is additive over items, one would have to administer 58 items to obtain the same precision about a typical examinee’s $\theta$ when using expected response curves, compared to using 39 items whose true parameters were known with certainty. This is a more honest estimate of the impact of using items whose parameters are known only through their modest relationships with available collateral information, to be weighed against the costs of obtaining information from a large calibration sample of examinees.

As mentioned above, the predictive distributions built in Part 1 can also be used as prior distributions to augment information from examinee response data. This was done with a modified version of BILOG, using responses from a new sample of 250 Form 8 examinees. Multivariate normal posterior distributions were obtained, with Bayes modal estimates as means and covariance matrices for each item that reflected the sum of precision from the collateral-information based prior and 250 examinee responses. 3PL approximations to expected response curves were again generated. Figures 14 and 15 are
the resulting TCC and TIC, and Figures 16 and 17 are the MLEs and standard errors for the same sample of 250 examinees used in Figures 10 and 13. The TCC and individual MLEs are now quite accurate, in the sense of agreeing with estimates obtained with item parameter estimates from the baseline sample. Posterior variances for examinees' \( \Theta \)s practically match those obtainable with baseline item parameter estimates.

By exploiting collateral information about items in a framework that appropriately accounts for the remaining uncertainty, it was possible in this example to obtain consistent estimates of examinee abilities and honestly state the uncertainty about them—with no response data at all for the items used to measure the examinees. Using the same collateral data to generate a prior distribution for item parameters, a supplemental calibration sample of 250 examinees provided estimates nearly indistinguishable from those obtained with the baseline item parameters with 5000 responses or more per item.

Conclusion

The title of this paper is a bit of a come-on; the techniques we describe don't really equate tests without any data at all. The point is, though, that the data they require are not the same pretesting- and equating-sample examinee data upon which previous equating procedures have traditionally relied. Years of research have shown that collateral information about items can be predictive of item operating characteristics. Recent developments in statistical methodologies make it possible to exploit this information in the equating problem, while giving an honest account of the consequences of the remaining uncertainties. There is no assurance that the collateral information about items available in any particular application will be sufficiently rich to eliminate or substantially reduce pretesting and equating. This remains to be discovered case by case. We now hope to explore the potential of the approach in a variety of settings.
References


Equating with Little or No Data

Page 21


Table 1
Descriptive Statistics and Parameter Estimates from Multivariate Regression Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation with Item Difficulty</th>
<th>Parameters in Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rater 1</td>
<td>Rater 2</td>
</tr>
<tr>
<td>The Item Passage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Syllable Words per 100 Words</td>
<td>.14</td>
<td>.20</td>
</tr>
<tr>
<td>Sentences per 100 Words</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>The Item Stem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Closed?</td>
<td>.11</td>
<td>.10</td>
</tr>
<tr>
<td>Hidden Negative?</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Line References?</td>
<td>.11</td>
<td>.11</td>
</tr>
<tr>
<td>The Options</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Arguments</td>
<td>.18</td>
<td>.26</td>
</tr>
<tr>
<td>Aspects of Targetted Solution Strategy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translate Active &amp; Passive</td>
<td>-.16</td>
<td>-.05</td>
</tr>
<tr>
<td>Translate Positive &amp; Negative</td>
<td>.04</td>
<td>.15</td>
</tr>
<tr>
<td>Process Single Sentence</td>
<td>-.08</td>
<td>-.18</td>
</tr>
<tr>
<td># Steps</td>
<td>.30</td>
<td>.20</td>
</tr>
<tr>
<td>Residual Covariance Matrix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>.05156</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>.01821</td>
<td>.49404</td>
</tr>
<tr>
<td>Lower Asymptote</td>
<td>-.00130</td>
<td>-.00161</td>
</tr>
</tbody>
</table>
List of Figures

1. Point Predictions of Item Parameters versus Baseline Estimates
2. Test Characteristic Curves from Point Predictions of Item Parameters and Baseline Estimates.
3. IRT True-Score Equating Curves based on Point Predictions of Item Parameters and Baseline Estimates
4. Examinee MLEs based on Point Predictions of Item Parameters and Baseline Estimates
5. Comparison of Examinee Standard Errors Calculated with Point Predictions of Item Parameters and Baseline Estimates in Place of True Item Parameters
6. The Effect of Uncertainty about \(b\) on Estimated Probabilities of Correct Response
7. Distribution for the Probability of a Correct Response at \(\theta=1\) Induced by the Uncertainty about \(b\)
8. Item Trace Lines Calculated with Baseline Estimates and Point Predictions of Item Parameters, and Parametric and Nonparametric Expected Response Curves
9. Test Characteristic Curves from Expected Response Curves and Baseline Estimates of Item Parameters
10. Examinee MLEs based on Expected Response Curves and Baseline Estimates of Item Parameters
11. IRT True-Score Equating Curves based on Expected Response Curves and Baseline Estimates of Item Parameters
12. Test Information Curves based on Expected Response Curves, and Point Predictions and Baseline Estimates of Item Parameters
13. Comparison of Examinee Standard Errors Calculated with Expected Response Curves and with Baseline Estimates of True Item Parameters
14. Test Characteristic Curves from Baseline Estimates of Item Parameters and Expected Response Curves based on Collateral Information and 250 Examinees
15. Test Information Curves based on Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees
16. Examinee MLEs based on Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees
17. Comparison of Examinee Standard Errors Calculated with Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees
FIGURE 1

Point Predictions of Item Parameters versus Baseline Estimates
FIGURE 2

Test Characteristic Curves from Point Predictions of Item Parameters and Baseline Estimates

--- BASELINE
--- PREDICTED
FIGURE 3

IRT True-Score Equating Curves based on Point Predictions of Item Parameters and Baseline Estimates
FIGURE 4

Examinee MLEs based on Point Predictions of Item Parameters and Baseline Estimates
FIGURE 5
Comparison of Examinee Standard Errors Calculated with Point Predictions of Item Parameters and Baseline Estimates in Place of True Item Parameters
The Effect of Uncertainty about $b$ on Estimated Probabilities of Correct Response

Figure 6
Distribution for the Probability of a Correct Response at $\theta = 1$

Induced by Uncertainty about $b$

Figure 7
Item Trace Lines Calculated with Baseline Estimates and Point Predictions of Item Parameters, and Parametric and Nonparametric Expected Response Curves
FIGURE 9

Test Characteristic Curves from Expected Response Curves and Baseline Estimates of Item Parameters
FIGURE 10
Examinee MLEs based on Expected Response Curves
and Baseline Estimates of Item Parameters
FIGURE 11

IRT True-Score Equiv. Curves based on Expected Response Curves and Baseline Estimates of Item Parameters
Test Information Curves based on Expected Response Curves, and Point Predictions and Baseline Estimate of Item Parameters
FIGURE 13
Comparison of Examinee Standard Errors Calculated with Expected Response Curves
and with Baseline Estimates of True Item Parameters
Figure 14
Test Characteristic Curves from Baseline Estimates of Item Parameters and Expected Response Curves based on Collateral Information and 250 Examinees
FIGURE 15

Test Information Curves based on Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees
Examinee MLEs based on Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees
Comparison of Examinee Standard Errors Calculated with Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees
<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr. Joseph Peterson</td>
<td>Naval Personnel Research and Development Center</td>
</tr>
<tr>
<td>Dr. Mark D. Rapone</td>
<td>ACT, F.O. Box 106, Iowa City, IA 52240</td>
</tr>
<tr>
<td>Dr. Frank L. Scharpf</td>
<td>Rice University, Texas 157, Houston, TX 77005</td>
</tr>
<tr>
<td>Dr. Robert Lusky</td>
<td>Texas A&amp;M University, College Station, TX 77845</td>
</tr>
<tr>
<td>Dr. W. V. Smith</td>
<td>University of California, Berkeley, CA 94720</td>
</tr>
<tr>
<td>Dr. Donald Rubin</td>
<td>Stanford University, Stanford, CA 94305</td>
</tr>
<tr>
<td>Dr. Robert Seeman</td>
<td>NASA Goddard Space Flight Center, Greenbelt, MD 20771</td>
</tr>
<tr>
<td>Dr. Victor L. I. Shilling</td>
<td>NASA MS 203, Code 22, Greenbelt, MD 20771</td>
</tr>
<tr>
<td>Dr. Kathleen Beamish</td>
<td>Naval Personnel Research and Development Center</td>
</tr>
<tr>
<td>Dr. Kurk Shumaker</td>
<td>NASA Ames Research Center, Moffett Field, CA 96035</td>
</tr>
<tr>
<td>Dr. Judy Spry</td>
<td>P.O. Box 470, Iowa City, IA 52243</td>
</tr>
<tr>
<td>Dr. Martin Seabing</td>
<td>Educational Testing Service, Princeton, NJ 08541</td>
</tr>
<tr>
<td>Dr. William Sexton</td>
<td>University of Illinois, Champaign, IL 61801</td>
</tr>
<tr>
<td>Dr. Kihun Tsubota</td>
<td>Educational Testing Service, Mail Stop 90-T, Princeton, NJ 08541</td>
</tr>
<tr>
<td>Dr. David Thomas</td>
<td>Psychometric Laboratory, CB# 377, Davis Hall, University of California,</td>
</tr>
<tr>
<td></td>
<td>Berkeley, CA 94704</td>
</tr>
<tr>
<td>Dr. Thomas J. Thomas</td>
<td>Federal Express Corporations, Human Resource Development, 1019 Director</td>
</tr>
<tr>
<td></td>
<td>Road, Suite 501, Memphis, TN 38131</td>
</tr>
<tr>
<td>Dr. Gary Thompson</td>
<td>University of Illinois, Educational Psychology, Champaign, IL 61801</td>
</tr>
<tr>
<td>Dr. Howard Wiemer</td>
<td>Educational Testing Service, Princeton, NJ 08541</td>
</tr>
<tr>
<td>Elizabeth Wald</td>
<td>Office of Naval Research, Code 377, 800 North Quincy Street, Arlington,</td>
</tr>
<tr>
<td></td>
<td>VA 22217-5000</td>
</tr>
<tr>
<td>Dr. Michael T. Wiemer</td>
<td>University of Wisconsin-Madison, Educational Psychology Dept., Box 413,</td>
</tr>
<tr>
<td></td>
<td>Madison, WI 53706</td>
</tr>
<tr>
<td>Dr. Ming-Hsi Wang</td>
<td>Educational Testing Service, Mail Stop 45-T, Princeton, NJ 08541</td>
</tr>
<tr>
<td>Dr. Thomas A. Wurm</td>
<td>FAA Academy, P.O. Box 20032, Oklahoma City, OK 73125</td>
</tr>
<tr>
<td>Dr. David J. Weiss</td>
<td>NASA Johnson Space Center, Houston, TX 77058</td>
</tr>
<tr>
<td>Dr. Douglas Winstead</td>
<td>Code 13, Naval Personnel Research and Development Center, San Diego, CA</td>
</tr>
<tr>
<td></td>
<td>92152-6000</td>
</tr>
<tr>
<td></td>
<td>German Military</td>
</tr>
<tr>
<td></td>
<td>Embassy, Post Office Box 361, D-5000 Enden 90, WEST GERMANY</td>
</tr>
</tbody>
</table>
Dr. David Wiley  
School of Education  
and Social Policy  
Northwestern University  
Evanston, IL 60208

Dr. Bruce Williams  
Department of Educational Psychology  
University of Illinois  
Urbana, IL 61801

Dr. Mark Wilson  
School of Education  
University of California  
Berkeley, CA 94720

Dr. Eugene Winnie  
Department of Psychology  
Emory University  
Atlanta, GA 30322

Dr. Martin F. Wade  
FERSERC  
99 Pacific St., Suite 4556  
Monterey, CA 93940

Mr. John H. Wolfe  
New Personnel R&D Center  
San Diego, CA 92122-4800

Dr. Kenji Yasumoto  
65-67  
Educational Testing Service  
Roseneate Road  
Princeton, NJ 08543

Ms. Doris Yen  
Educational Testing Service  
Princeton, NJ 08543

Dr. Wendy Yen  
CTB/McGraw Hill  
Del Monte Research Park  
Monterey, CA 93940

Dr. Jaemee L. Young  
National Science Foundation  
Room 320  
1800 G Street, N.W.  
Washington, DC 20550