RR-92-20-ONR

2

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE February, 1992	3. REPORT TYPE AND DATES COVERED Final

4. TITLE AND SUBTITLE

How to Equate Tests with Little or No Data

5. FUNDING NUMBERS

G. N00014-88-K-0304
PE. 61153N
PR. RR 04204
TA. RR 04204-01
WU. R&T 4421552

6. AUTHOR(S)

Robert J. Mislevy, Kathleen M. Sheehan &
Marilyn Wingersky

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Educational Testing Service
Rosedale Road
Princeton, NJ 08541

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Cognitive Sciences
Code 1142CS
Office of Naval Research
Arlington, VA 22217-5000

10. SPONSORING MONITORING AGENCY REPORT NUMBER

N/A

11. SUPPLEMENTARY NOTES

None

12a. DISTRIBUTION AVAILABILITY STATEMENT

Unclassified/Unlimited

12b. DISTRIBUTION CODE

N/A

13. ABSTRACT (Maximum 200 words)

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).

14. SUBJECT TERMS

Bayesian estimation, cognitive processes, collateral information, equating, item response theory

15. NUMBER OF PAGES

46 + RDP

16. PRICE CODE

N/A

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

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 Nueral 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.[1] 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 orderngs 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 tc high relationships between item features and operating characteristics are of conside¬able 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 samples (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:

$$F_j(\theta) \equiv P(X_j=1|\theta,\beta_j)\,, \tag{1}$$

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

$$F_j(\theta) \equiv c_j + (1-c_j)\,\Psi\left[a_j(\theta-b_j)\right];$$

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

$$p(\mathbf{x}|\theta,\mathbf{B}) = \prod_{j=1}^{n} F_j(\theta)^{x_j}\left[1-F_j(\theta)\right]^{1-x_j}\,, \tag{2}$$

where $\mathbf{B}=(\beta_1,\ldots,\beta_n)$.

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 θ 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 θ, $L(\theta|\mathbf{x},\mathbf{B})$, once **x** has been observed. The value of θ that maximizes L is the maximum likelihood estimate (MLE) $\hat{\theta}$. Its variance, $Var(\hat{\theta}|\theta,\mathbf{B})$, can be approximated by the second derivative of log L evaluated at $\hat{\theta}$. The posterior density of θ with respect to the prior density p(θ) is obtained as

$$p(\theta|\mathbf{x},\mathbf{B}) \propto L(\theta|\mathbf{x},\mathbf{B})\ p(\theta)\ . \tag{3}$$

The mean of (3) is the Bayes mean estimate $\bar{\theta}$; the variance, $Var(\theta|\mathbf{x},\mathbf{B})$, 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 θ is given by

$$\tau_A(\theta) = \sum_{j\in S_A} p(x_j=1|\theta,\beta_j) = \sum_{j\in S_A} F_j(\theta)\ , \tag{4}$$

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 θ, 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 θ, over the uncertainty associated with the observational setting. Specifically, the uncertainty in scores for a given θ in standard IRT true-score equating is the 0 or 1 for each x_j, with β_j assumed known.

[2] Under the 3PL, this relationship does not give equatings for scores below the sum of the c_js 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 **B** (e.g., Mislevy, 1986; Tsutakawa & Lin, 1986) begins with a (possibly uninformative) prior distribution p(**B**), a known or concurrently estimated examinee population density p(θ), and a response matrix $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ from a sample of N independently-responding examinees.[3] The posterior distribution of **B** is

$$p(\mathbf{B}|\mathbf{X}) \propto p(\mathbf{B})\ L(\mathbf{B}|\mathbf{X})\ , \qquad (5)$$

where $L(\mathbf{B}|\mathbf{X})$ is the marginal likelihood function for the item parameters (Bock & Aitkin, 1981):

$$L(\mathbf{B}|\mathbf{X}) = \prod_{i=1}^{N} \int p\left(\mathbf{x}_i|\theta_i,\mathbf{B}\right) p(\theta_i)\ d\theta_i\ . \qquad (6)$$

One can obtain Bayes mean estimates $\overline{\mathbf{B}}$ or Bayes modal estimates $\widetilde{\mathbf{B}}$, and a posterior variance matrix $\Sigma_{\mathbf{B}}$ from (5), leading to the approximations $p(\mathbf{B}|\mathbf{X}) \sim N(\overline{\mathbf{B}},\Sigma_{\mathbf{B}})$ or $N(\widetilde{\mathbf{B}},\Sigma_{\mathbf{B}})$. Alternatively, one obtains the MLE $\hat{\mathbf{B}}$ by maximizing (6) with respect to **B**. The consistency of $\overline{\mathbf{B}}$, $\widetilde{\mathbf{B}}$, and $\hat{\mathbf{B}}$ as estimators of **B** 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 $L(\theta|\mathbf{x},\mathbf{B}=\overline{\mathbf{B}})$ for $L(\theta|\mathbf{x},\mathbf{B})$ when estimating θ, or $p(x_j=1|\theta,\mathbf{B}=\hat{\mathbf{B}})$ for $p(x_j=1|\theta,\mathbf{B})$ when calculating $\tau_A(\theta)$ and $\tau_B(\theta)$ in equating (Lord, 1982).

If **B** is *not* well determined—i.e., p(**B**|“data relevant to **B**”) 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 θ. 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 (y_j) about a test item, and suggests ways to express this information in terms of distributions for the item parameters β_j. 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 $p(\beta|y_j)$ 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 into 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 β_j is the difficulty parameter of Item j; η_k is the contribution of Feature k to item difficulty, for k=1,...,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,...,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 ϕ^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

$$p(\beta_j|y_j) \approx N\left(\sum_{k=1}^{K} y_{kj}\hat{\eta}_k, \hat{\phi}^2\right),$$

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 β_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 $p(\mathbf{B}|\mathbf{X},\mathbf{Y})$.

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 $\beta_j \equiv (a_j, -(b_j/a_j), c_j)$, 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.

[Insert Table 1 about here]

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 θ. 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 discrepency 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 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 θ 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 $\hat{\theta}$s. A bias corresponding to the discrepencies 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 $\mathrm{Var}(\hat{\theta}|\theta,\mathbf{B})$, 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 θ estimates should be larger; simply calculating $\mathrm{Var}(\hat{\theta}|\theta,\mathbf{B})$ with $\overline{\mathbf{B}}$ in place of **B** neglects uncertainty about θs due to the remaining uncertainty about item parameters. We shall see that ignoring this uncertainty causes posterior variances for θs 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 $\bar{\beta}_j$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 β_j values. Biases in $\hat{\theta}$s appear because the $\bar{\beta}_j$s are too clustered around their average. More seriously, disregarding the uncertainty about item parameters causes substantial understatement of the uncertainty about θs. 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 θ with imperfect knowledge about **B**, conveyed through p(**B**|data), where "data" refers to a calibration-sample **X** of responses from N examinees, collateral information about items, or both. The probative value about θ 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:

$$L(\theta|\mathbf{x},\text{data concerning }\mathbf{B}) = \int L(\theta|\mathbf{x},\mathbf{B})\; p(\mathbf{B}|\text{data concerning }\mathbf{B})\; d\mathbf{B}\ . \tag{7}$$

Tsutakawa compared Bayesian inferences about θ using p(**B**|**X**) and $\mathbf{B}=\overline{\mathbf{B}}$, under the 2- and 3-parameter logistic models (the 2PL and 3PL). Under the 2PL, the more accurate estimates of Var(θ|**x**) using p(**B**|**X**) were higher than the usual approximation, $\text{Var}(\theta|\mathbf{x},\mathbf{B}=\overline{\mathbf{B}})$, 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 θ, 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**|data) in (4) to obtain expected scores:

$$\tau^{*}_{A}(\theta) \equiv E_{\mathbf{B}}\left[\tau_{A}(\theta)\right] = \sum_{j\in S_A} \int p(x_j=1|\theta,\beta_j)\; p(\beta_j|\text{data})\; d\beta_j\ . \tag{8}$$

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 θ and $\mathbf{B}$ is $F_j(\theta) \equiv P(x_j=1|\theta,\beta_j)$. If β_j is only partially known, through $p(\beta_j|\text{data})$, the probability of a correct response conditional on θ but marginal with respect to $\mathbf{B}$ can be written as

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

an "expected response curve" that gives the probability of correct response conditional on θ taking into account uncertainty about β_j (Lewis, 1985).

Even though $F_j^*(\theta)$ is the expected value of a correct response at each value of θ, it is *not* the same as $F_j(\theta)$ evaluated with the expected value of β_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 $p(b_j|\text{data})$ is $N(\bar{b}_j,\sigma_j^2)$. Then F_j^* is also 2PN, but with $b_j^*=\bar{b}_j$ and

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

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 $p(b) \sim N(0,1)$. 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 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 $F^*(1)$.

[Insert Figures 6 and 7 about here]

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

$$p(\mathbf{x}|\theta,\text{data concerning }\mathbf{B}) = \prod_{j=1}^{n} F_j^*(\theta)^{x_j}\,[1-F_j^*(\theta)]^{1-x_j}\,.. \qquad (9)$$

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

Equation (9) proves useful even if $p(\mathbf{B})$ 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 θ. Thus for large n, Bayesian and likelihood point estimates of θ based on (9) have the correct expectation. Indicators of their uncertainty based on (9), however, such as the variance estimator of $\hat{\theta}$ 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 θ from this source is minor.

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

$$\tau_A^*(\theta) = \sum_j F_j^*(\theta)\ . \tag{10}$$

Since only expectations are involved, (10) is correct whether or not $p(\mathbf{B})$ 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 θ values across the range of interest. Denote by Θ_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 Θ_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:

$$F_j^*(\Theta_m) \approx S^{-1}\sum_{s=1}^{S} P_{jm}^{(s)}\ .$$

Steps 2 and 3 generate an empirical approximation of the predictive distribution of $P(X_j=1|\theta,\beta_j)$ over the range of β_j for fixed values of θ, an example of which appeared as Figure 7. Step 4 is finding the posterior mean for P with respect to β_j conditional on each of the θ points—approximations of the values on the expected response curve. Subsequent inferences about θ can be drawn using these values directly in a discrete approximation of integrals involving θ 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 β^{**} 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 θ scale. In practical work, one might create simulated examinees at each Θ_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 θ" 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 θ grid from -3 to +3 in steps of .2, the average modelled percent-correct was evaluated from each of the 100 plausible values of β_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 preceeding 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 θs 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 θ 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.

[Insert Figures 9-13 about here]

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 are 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' θs practically match those obtainable with baseline item parameter estimates.

[Insert Figures 14-17 about here]

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

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

Arnold, B.C., & Strauss, D. (1988). Pseudolikelihood estimation. *Technical Report No. 164*. Riverside, CA: Department of Statistics, University of California.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303-310.

Bejar, I.I. (1985). Speculations on the future of test design. In S.E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 279-294). Orlando: Academic Press.

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46*, 443-459.

Chaffin, R., & Peirce, L. (1988). A taxonomy of semantic relations for the classification of GRE analogy items. *Research Report RR-87-50*. Princeton, NJ: Educational Testing Service.

Chalifour, C., & Powers, D.E. (1989). The reationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement, 26*, 120-132.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Dorans, N. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3*, 3-17.

Drum, P.A., Calfee, R.C., & Cook, L.K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly, 16*, 486-514.

Embretson, S.E. (Ed.) (1985). *Test design: Developments in psychology and psychometrics*. Orlando: Academic Press.

Enright, M.K., & Bejar, I.I. (1989). An analysis of test writers' expertise: Modeling analogy item difficulty. *Research Report RR-89-35*. Princeton, NJ: Educational Testing Service.

Fischer, G.H. (1973). The linear logistic test model as an instrument of educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G.H., & Formann, A.K. (1982). Some applications of the logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6*, 397-416.

Guttman, L. (1959). A structural theory for inter-group beliefs and actior.. *American Sociological Review, 24*, 318-328.

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 of Education/Macmillan.

Hively, W., Patterson, H.L., & Page, S.H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

Irvine, S.H., Dann, P.L., & Anderson, J.D. (in press). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology.*

Lewis, C. (1985). Estimating individual abilities with imperfectly known item response functions. Paper presented at the Annual Meeting of the Psychometric Society, Nashville TN, June, 1985.

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

Lord, F.M. (1982). Item response theory and equating—A technical summary. In P.W. Holland & D.B. Rubin (Eds.), *Test equating* (pp. 141-148). New York: Academic Press.

Lorge, I, & Kruglov, L. (1952). A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological M:easurement, 12*, 554-561.

Lorge, I., & Kruglov, L. (1953). The improvement of estimates of test difficulty. *Educational and Psychological Measurement, 13*, 34-46.

Mayer, R.E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 10*, 135-175.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-196.

Mislevy, R.J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12*, 281-296.

Mislevy, R.J., & Bock, R.D. (1983). *BILOG: Item analysis and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific software, Inc.

Mislevy, R.J., & Sheehan, K.M. (1989a). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661-679.

Mislevy, R.J., & Sheehan, K.M. (1989b). Information matrices in latent-variable models. *Journal of Educational Statistics, 14*, 335-350.

Mislevy, R.J., & Yan, D. (in press). Dealing with uncertainty about item parameters: Multiple imputations and SIR. RR-92-xx-ONR. Princeton: Educational Testing Service.)

Mitchell, K.J. (1983). Cognitive processing determinants of item difficulty on the verbal subtests of the Armed Services Vocational Aptitude Battery. *Technical Report 598*. Alexandria, VA: U.S. Army Reseaich Institute for the Behavioral and Social Sciences.

Osburn, H.G. (1968). Item sampling for achievemen: :esting. *Educational and Psychological Measurement, 28*, 95-104.

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

Scheiblechner, H. (1972). Das lernen und lösen komplexer denkaufgaben. *Zeitschrift für Experimentelle und Angewandte Psychologie, 19*, 476-506.

Scheuneman, J., Gerritz, K., & Embretson, S. (1989). Effects of prose complexity on achievement test item difficulty. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, March 1989.

Tatsuoka, K.K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement, 24*, 233-245.

Thorndike, R.L. (1982). Item and score conversion by pooled judgment. In P.W. Holland & D.B. Rubin (Eds.), *Test equating* (pp. 309-326). New York: Academic Press.

Tinkelman, S. (1947). Difficulty prediction of test items. *Teachers College Contributions to Education*, No. 941. New York: Teachers College, Columbia university.

Tsutakawa, R.K., & Johnson, J. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371-390.

Tsutakawa, R.K., & Lin, H.Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51*, 251-267.

Tsutakawa,R.K., & Soltys, M.J. (1988). Approximation for Bayesian ability estimation. *Journal of Educational Statistics, 13*, 117-130.

Whitely, S.E. (1976). Solving verbal analogies: Some cognitive components of intelligence test items. *Journal of Educational Psychology, 68*, 234-242.

Wingersky, M.S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R.K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, B.C.: Educational Research Institute of British Columbia.

Table 1

Descriptive Statistics and Parameter Estimates from Multivariate Regression Model

	Correlation with Item Difficulty			Parameters in Regression Model		
Variable	Rater 1	Rater 2	% Rater Agreement	Slope	Intercept	Lower Asymptote
The Item Passage						
3 Syllable Words per 100 Words	.14	.20	.91		-.02321	
Sentences per 100 Words	.01	.01	.93		.11101	
The Item Stem						
Closed?	.11	.10	.99		-.19720	
Hidden Negative?	.00	.00	.99			-.16061
Line References?	.11	.11	.96		-.48298	
The Options						
# Arguments	.18	.26	.93		-.07365	-.00190
Aspects of Targetted Solution Strategy						
Translate Active & Passive	-.16	-.05	.90	.19295	.36407	
Translate Positive & Negative	.04	.15	.95		-.74103	
Process Single Sentence	-.08	-.18	.83		.12783	
# Steps	.30	.20	.70		-.11304	
Residual Covariance Matrix						
Slope				.05156		
Intercept				.01821	.49404	
Lower Asymptote				-.00130	-.00161	.00121

List of Figures

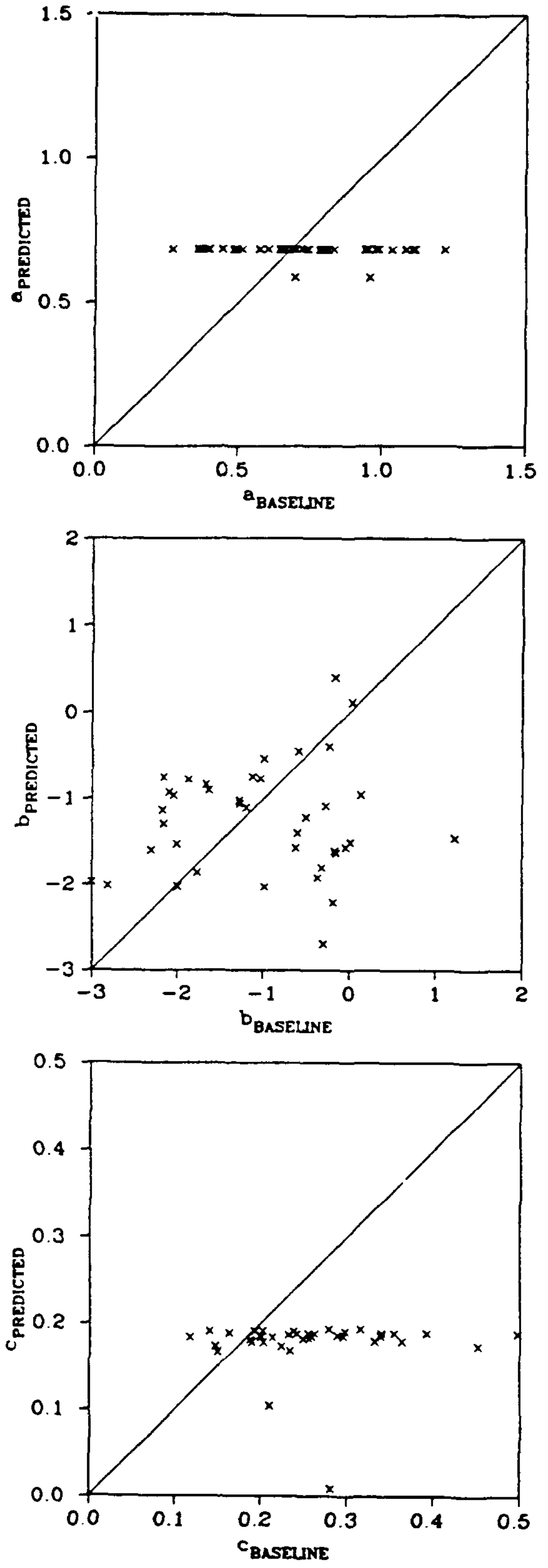


FIGURE 1

Point Predictions of Item Parameters versus Baseline Estimates

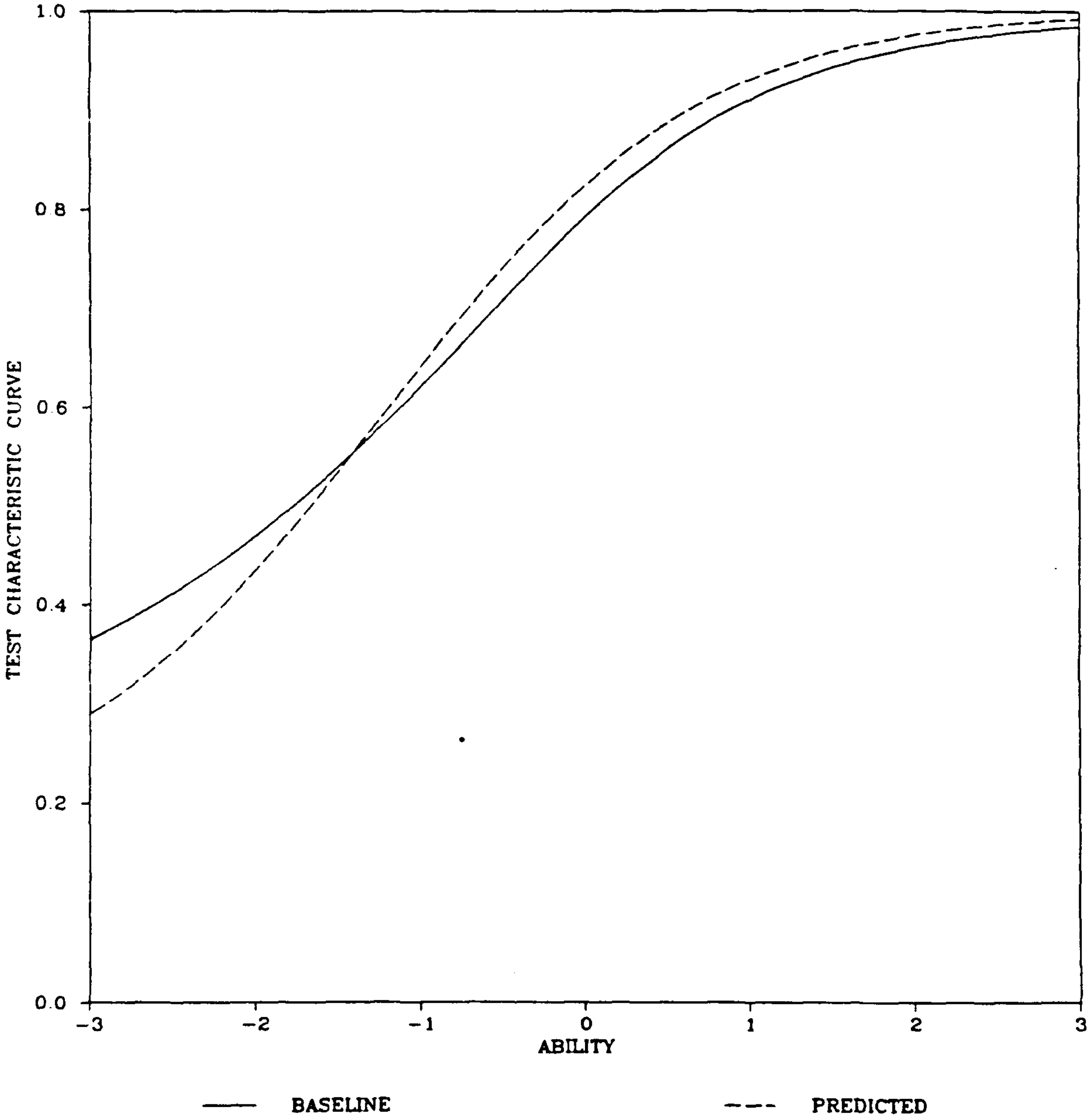


FIGURE 2

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

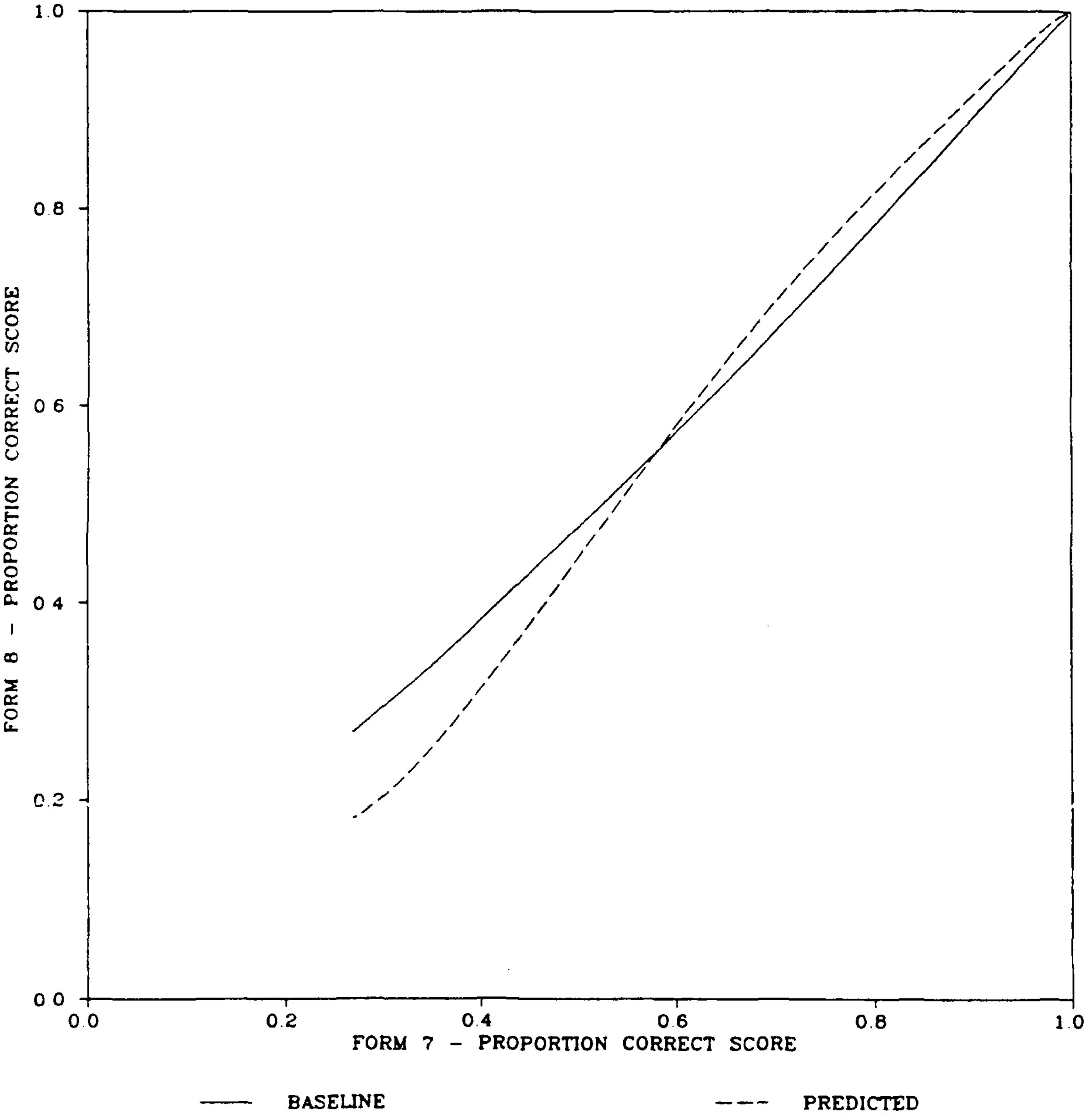


FIGURE 3

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

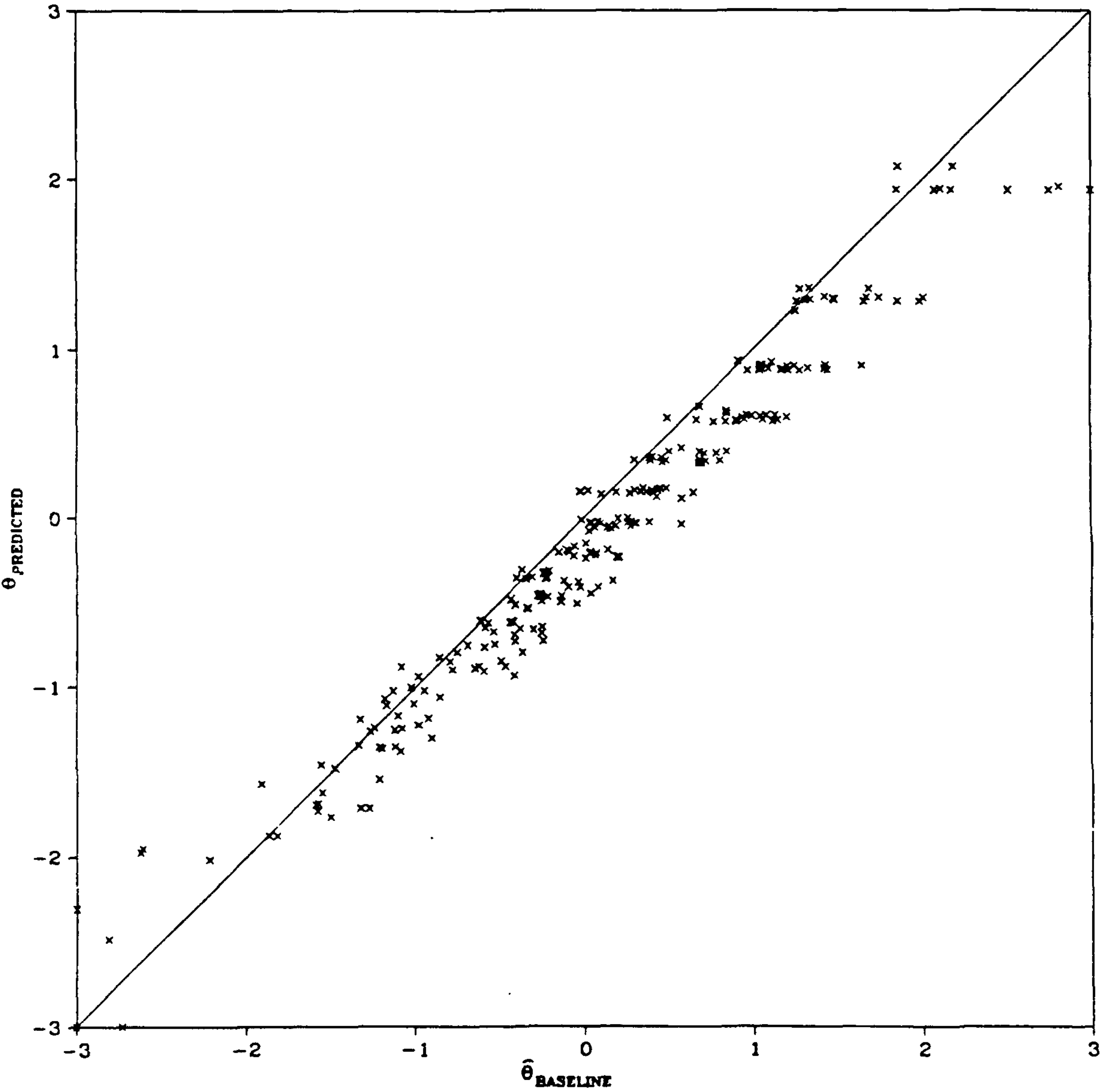


FIGURE 4

Examinee MLEs based on Point Predictions of
Item Parameters and Baseline Estimates

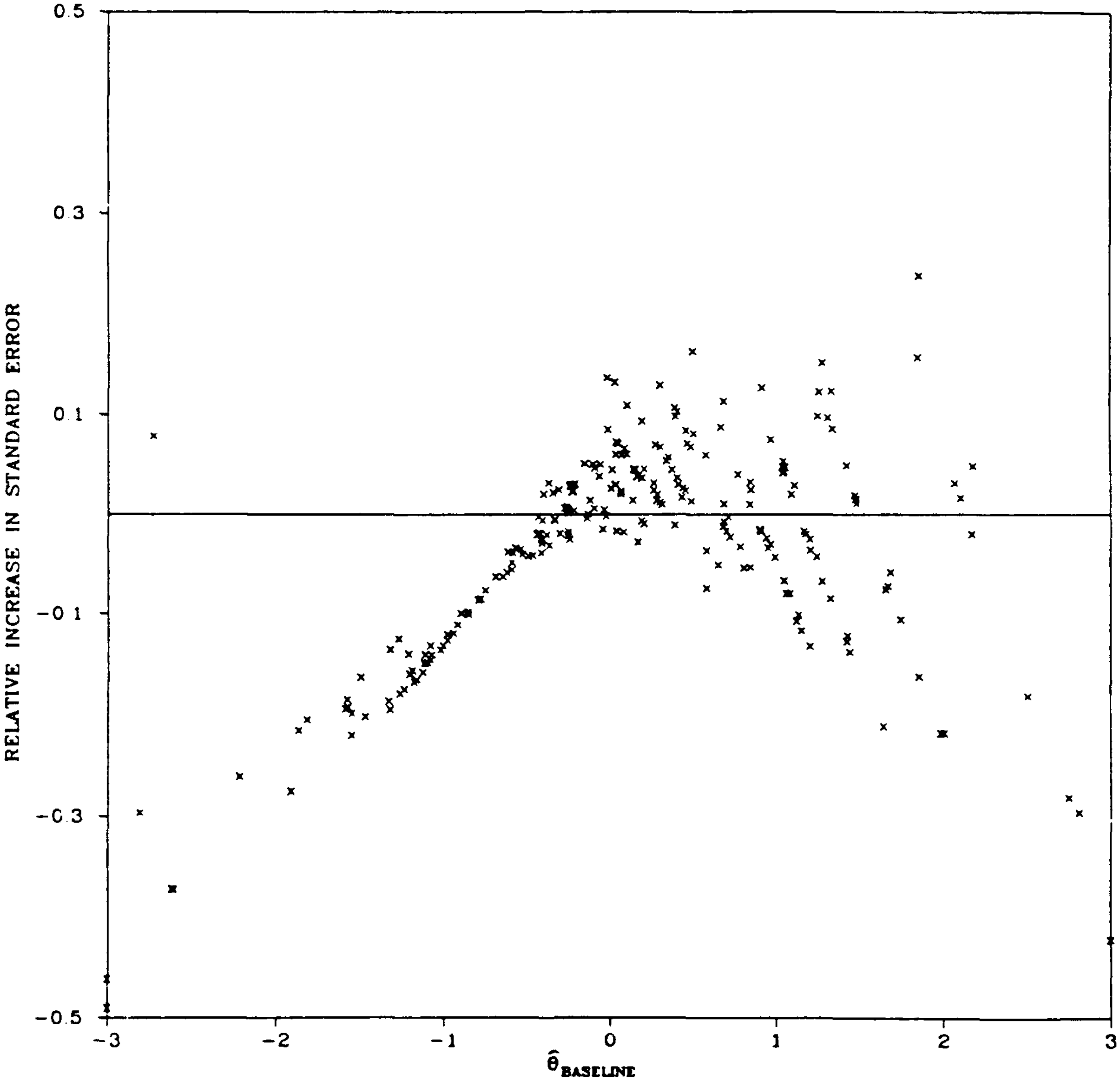
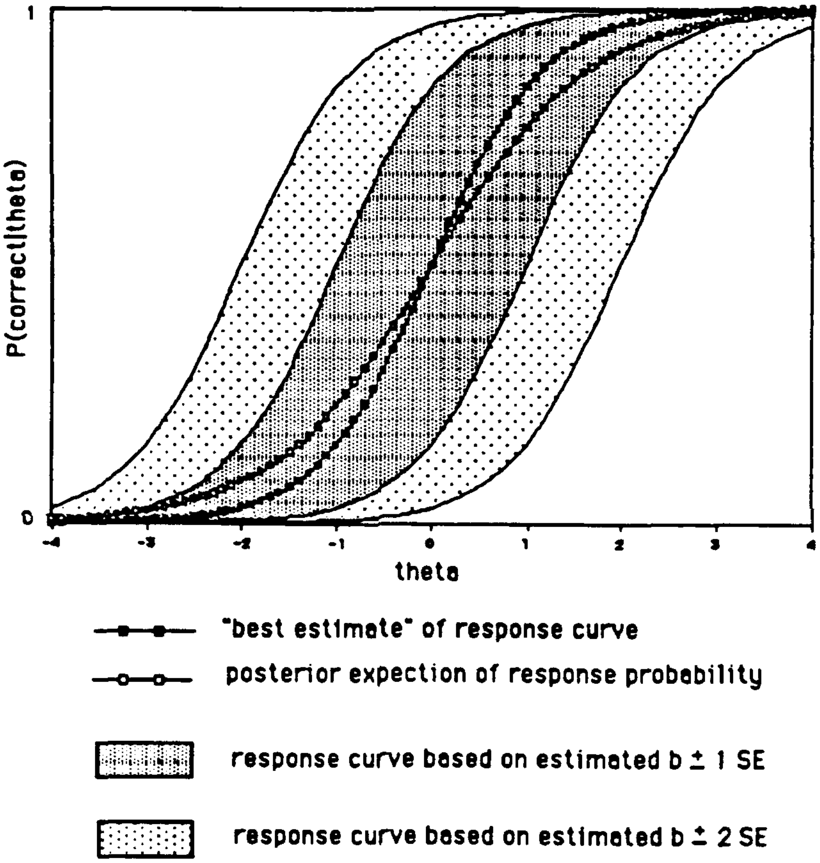


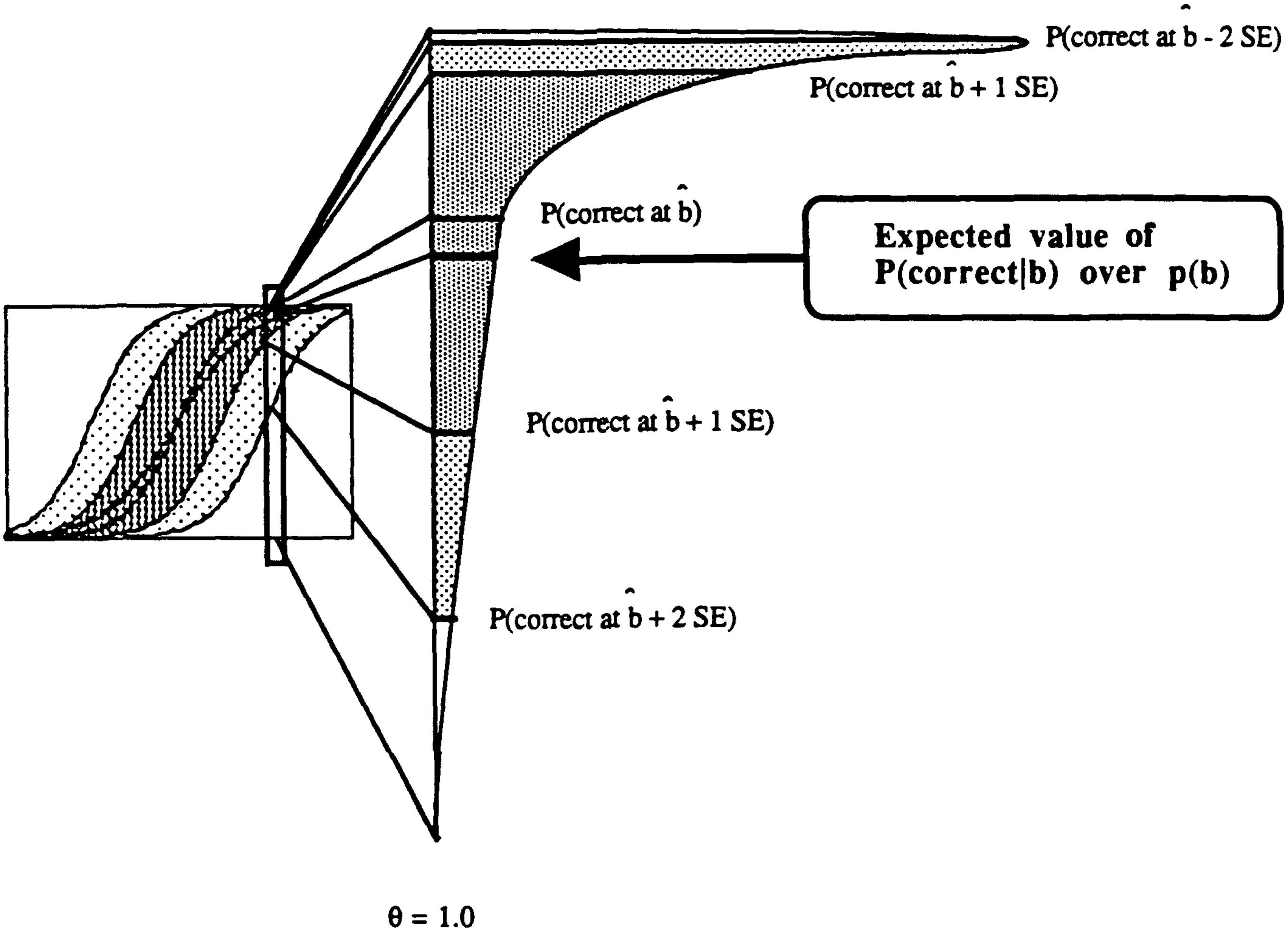
FIGURE 5

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


P(correct at $\hat{b}$ - 2 SE)
P(correct at $\hat{b}$ + 1 SE)
P(correct at $\hat{b}$)
Expected value of P(correct|b) over p(b)
P(correct at $\hat{b}$ + 1 SE)
P(correct at $\hat{b}$ + 2 SE)
θ = 1.0

Distribution for the Probability of a Correct Response at θ=1

Induced by Uncertainty about b

Figure 7

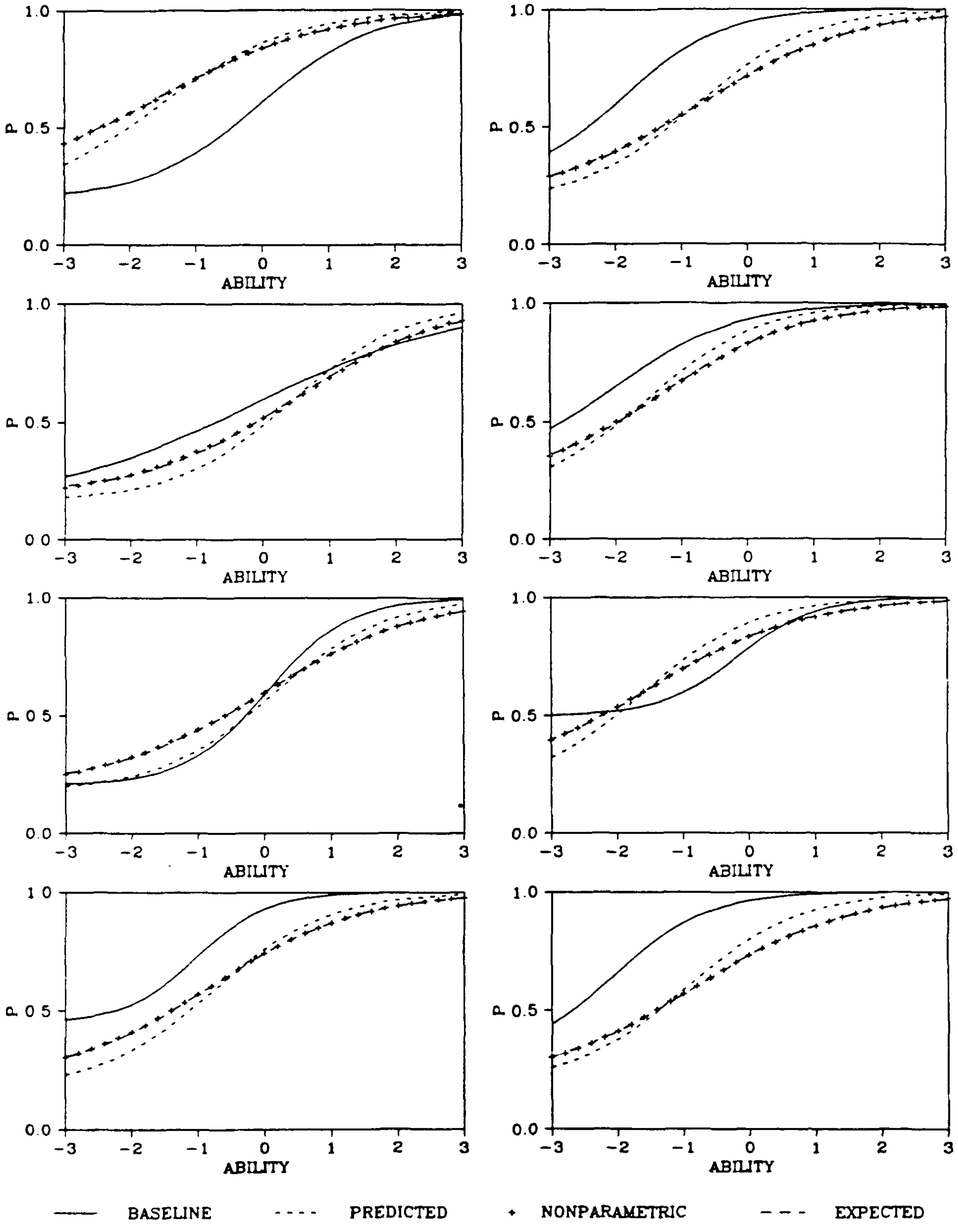


FIGURE 8

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

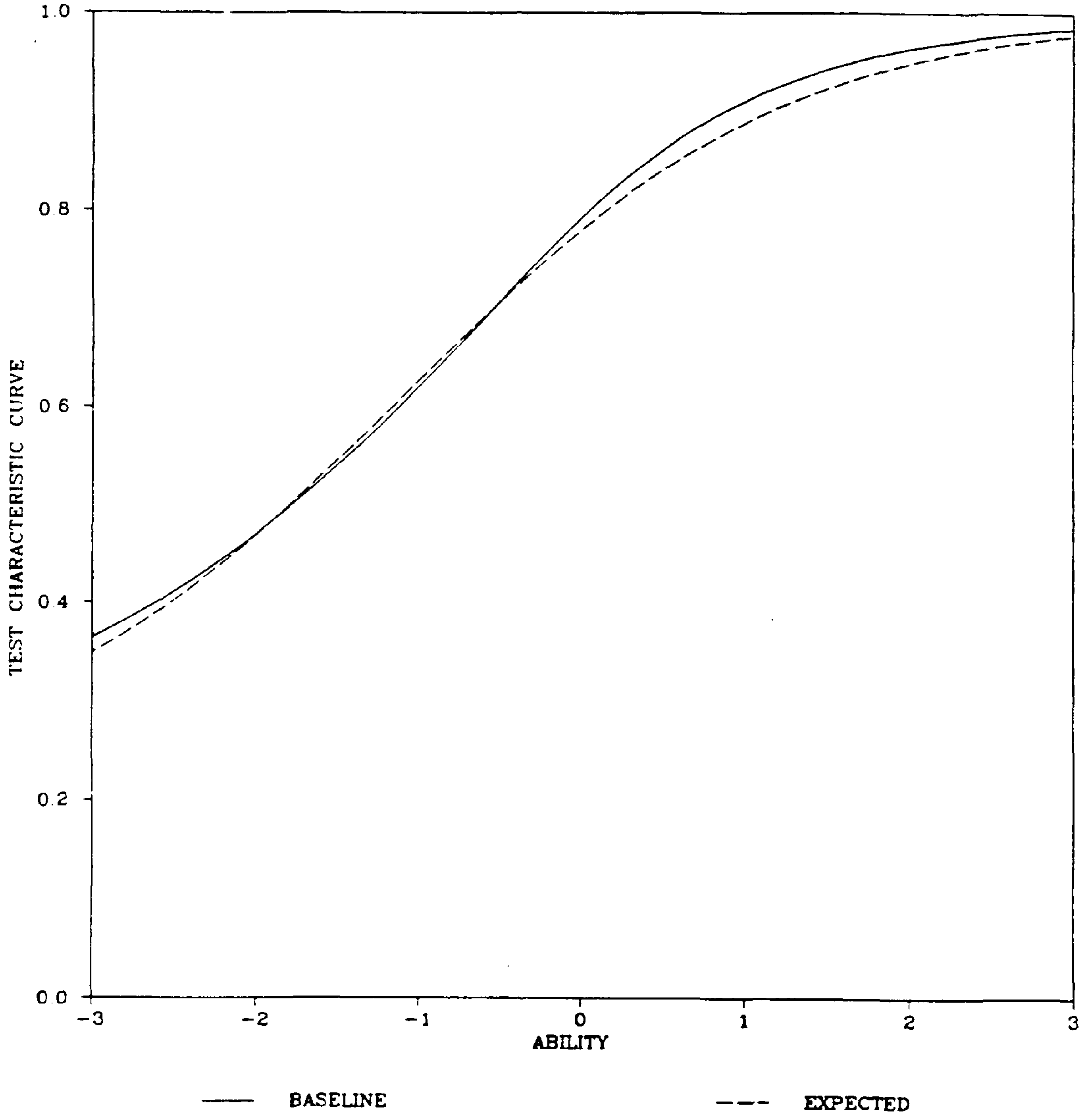


FIGURE 9

Test Characteristic Curves from Expected Response Curves and Baseline Estimates of Item Parameters

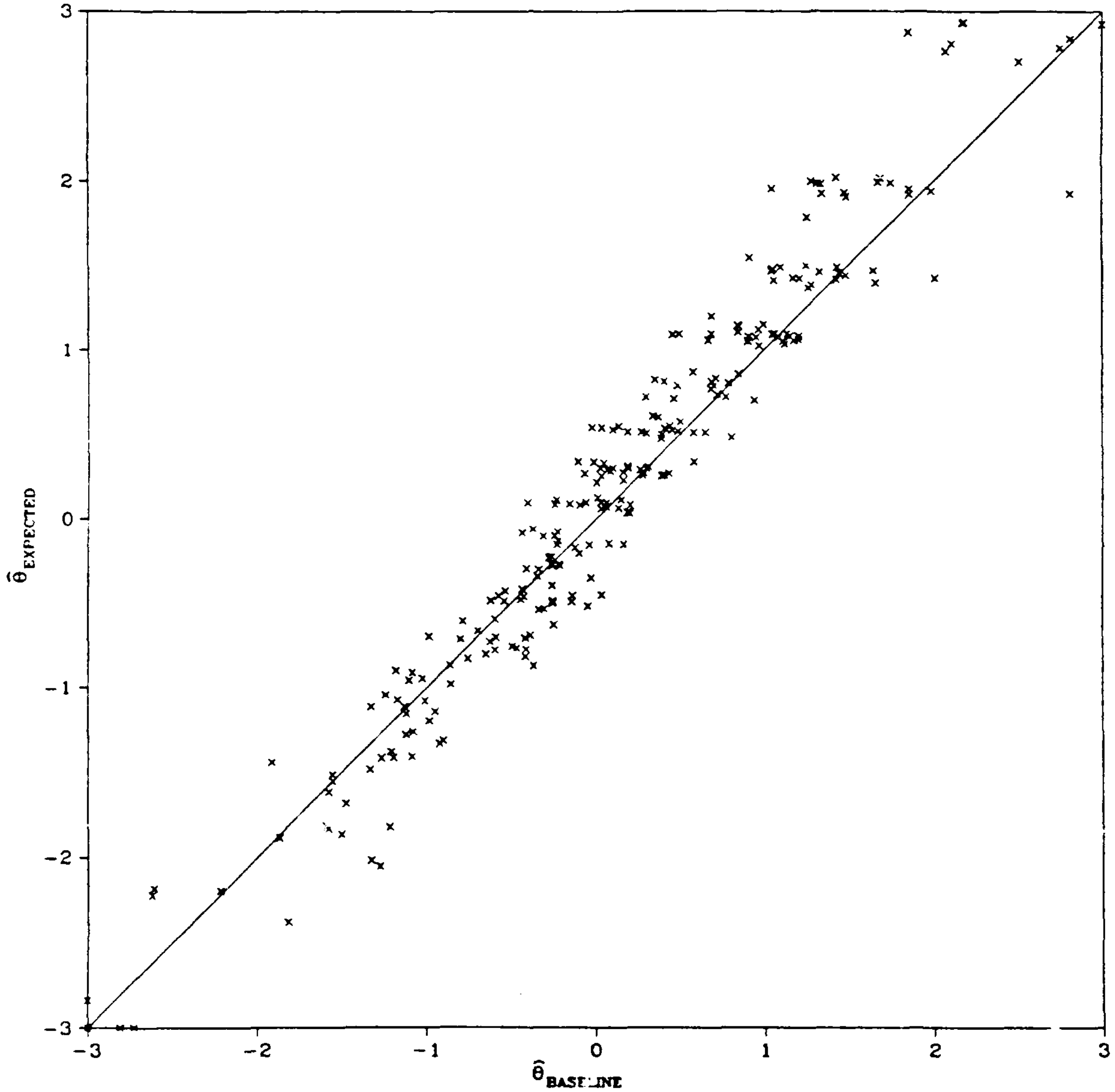


FIGURE 10

Examinee MLEs based on Expected Response Curves and Baseline Estimates of Item Parameters

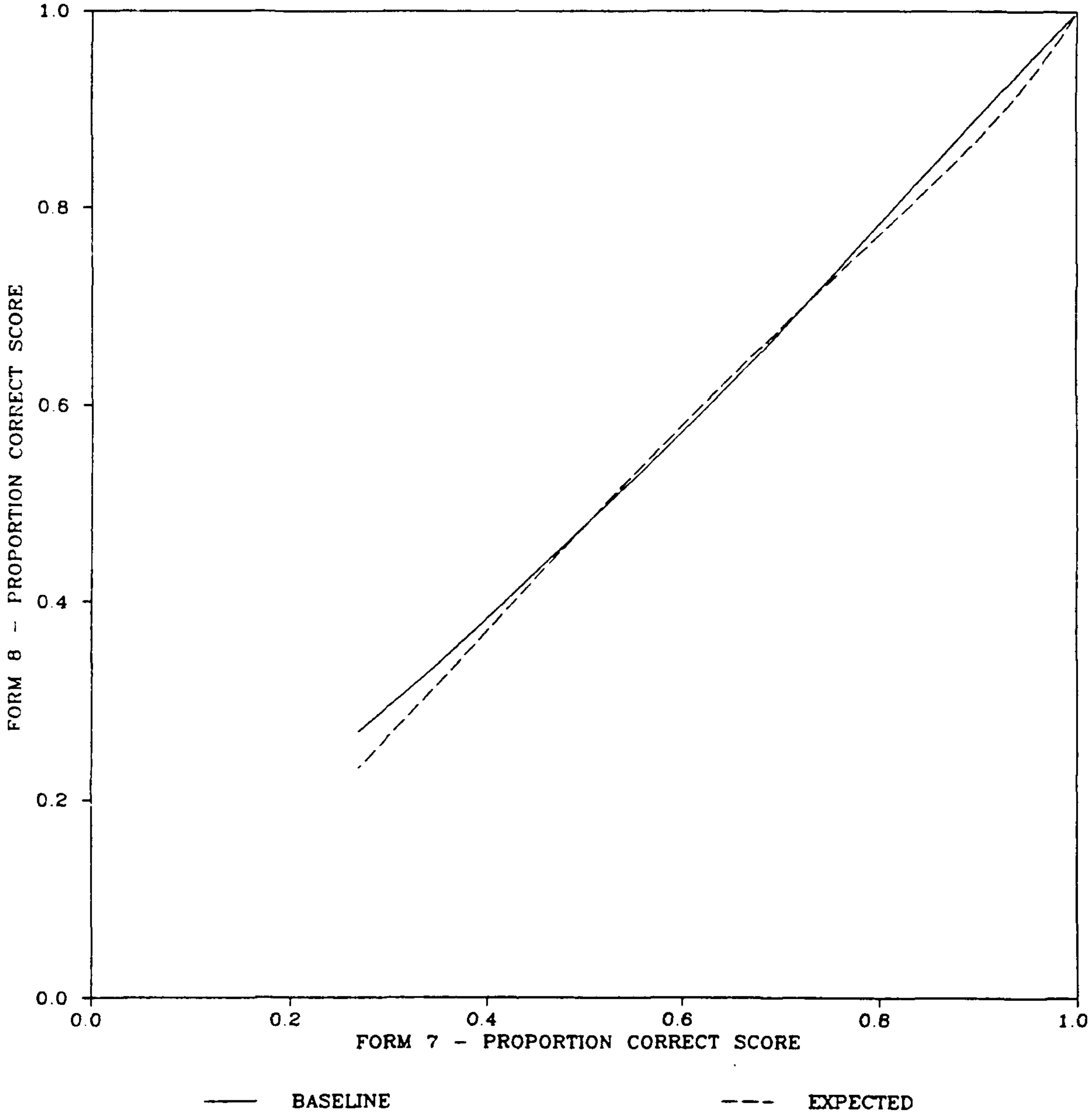


FIGURE 11

IRT True-Score Equa , Curves based on Expected Response Curves and Baseline Estimates of Item Parameters

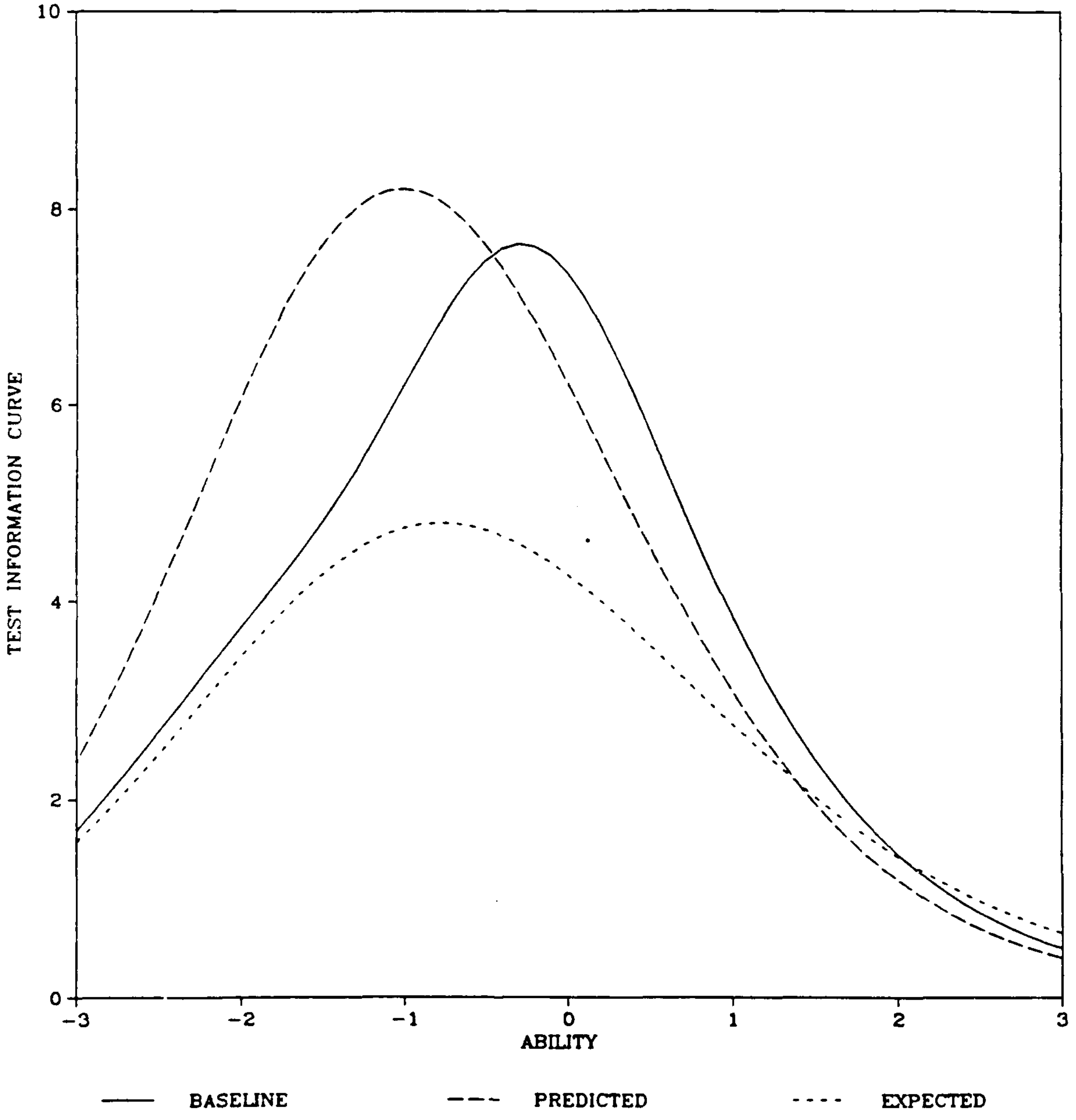


FIGURE 12

Test Information Curves based on Expected Response Curves, and Point Predictions and Baseline Estimate: of Item Parameters

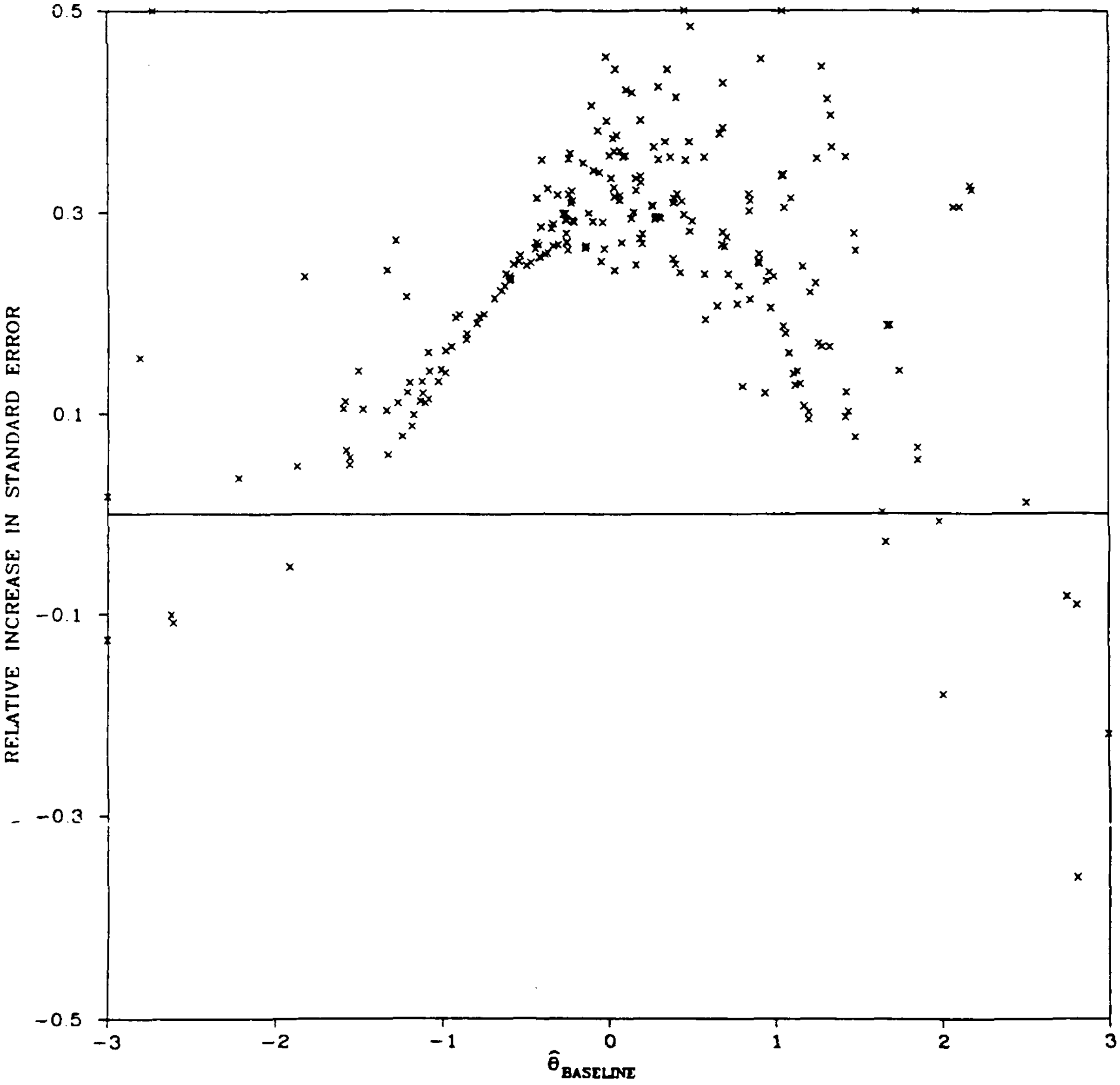


FIGURE 13

Comparison of Examinee Standard Errors Calculated with Expected Response Curves and with Baseline Estimates of True Item Parameters

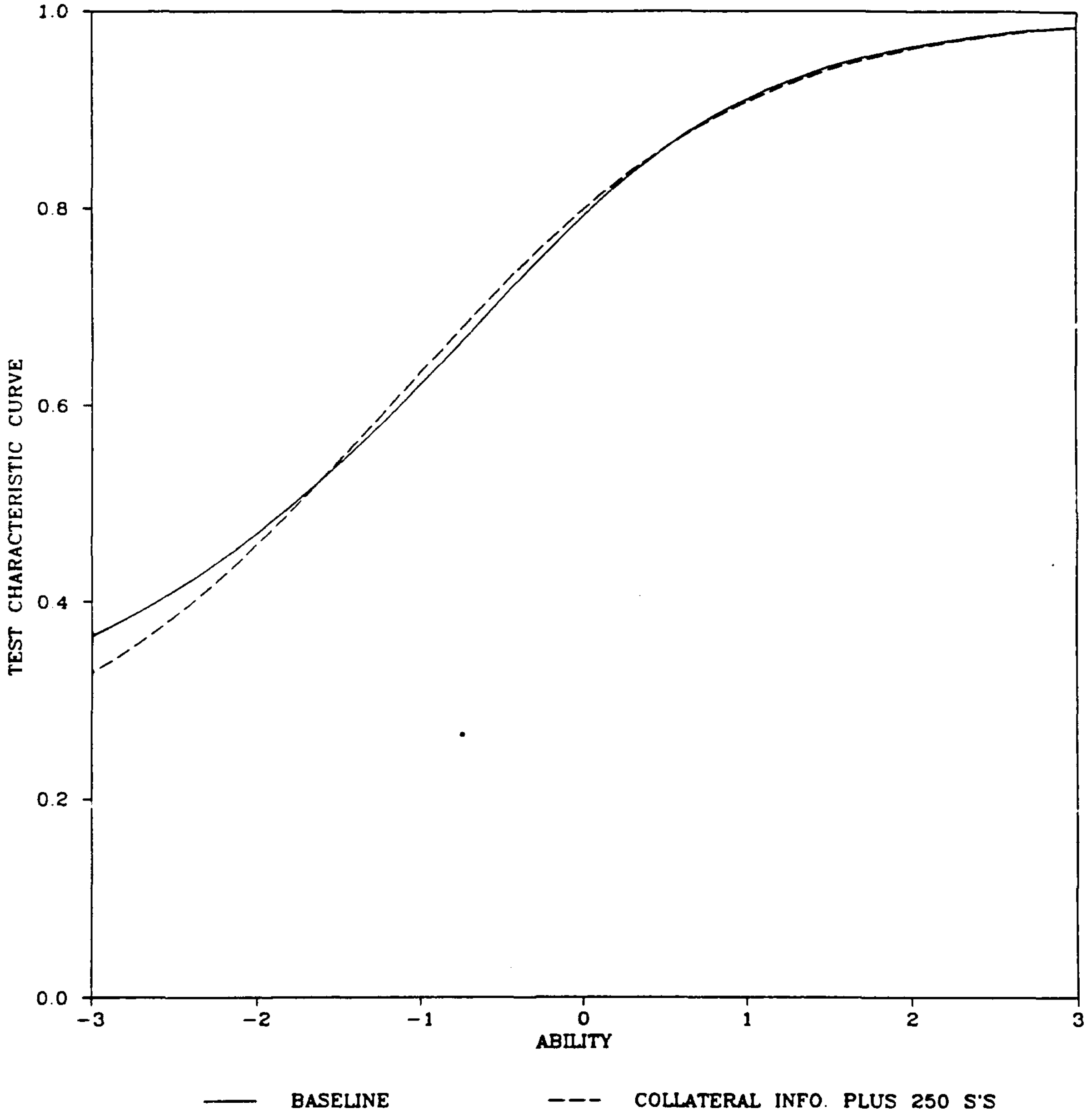


FIGURE 14

Test Characteristic Curves from Baseline Estimates of Item Parameters and Expected Response Curves based on Collateral Information and 250 Examinees

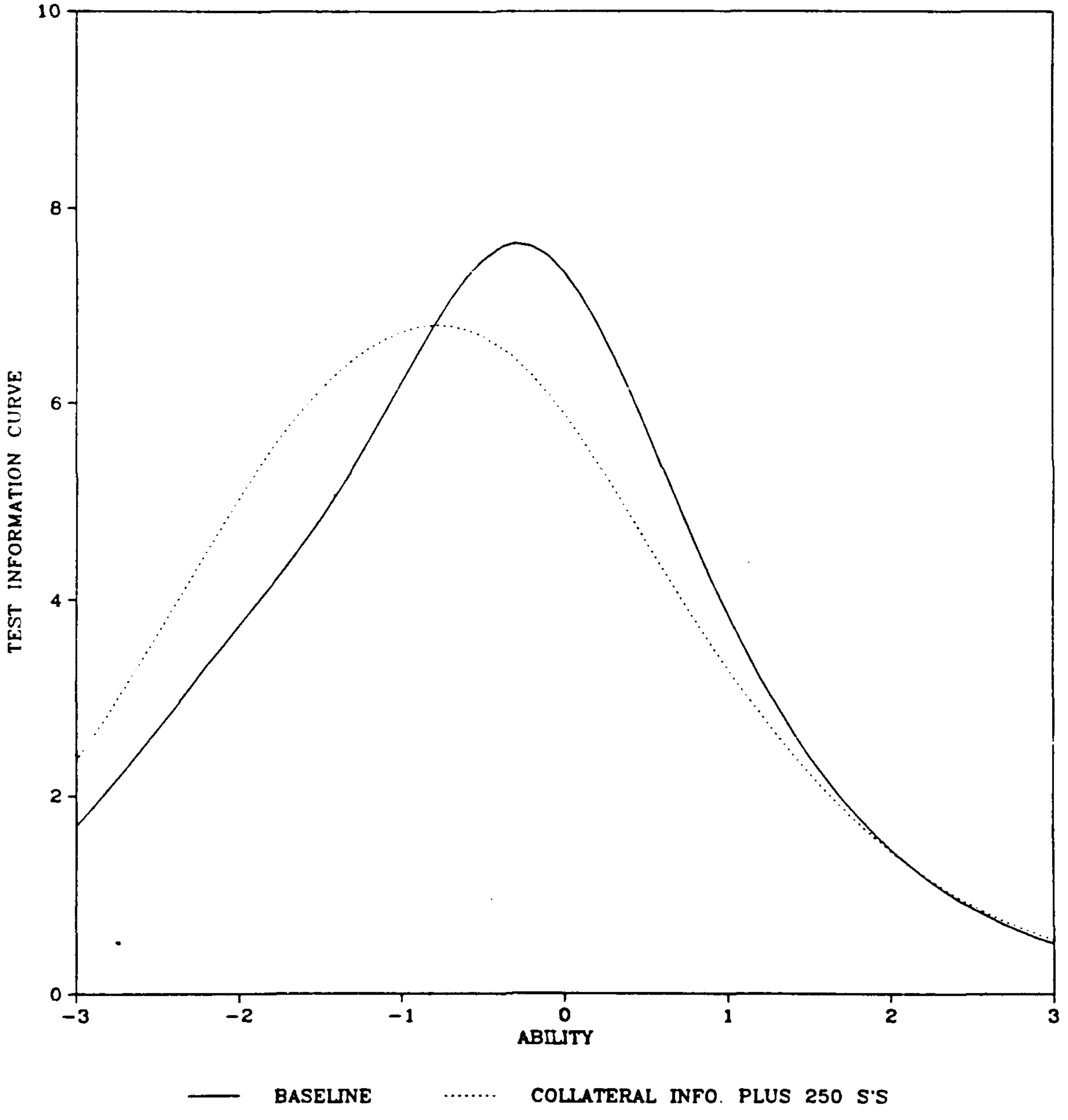


FIGURE 15

Test Information Curves based on Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees

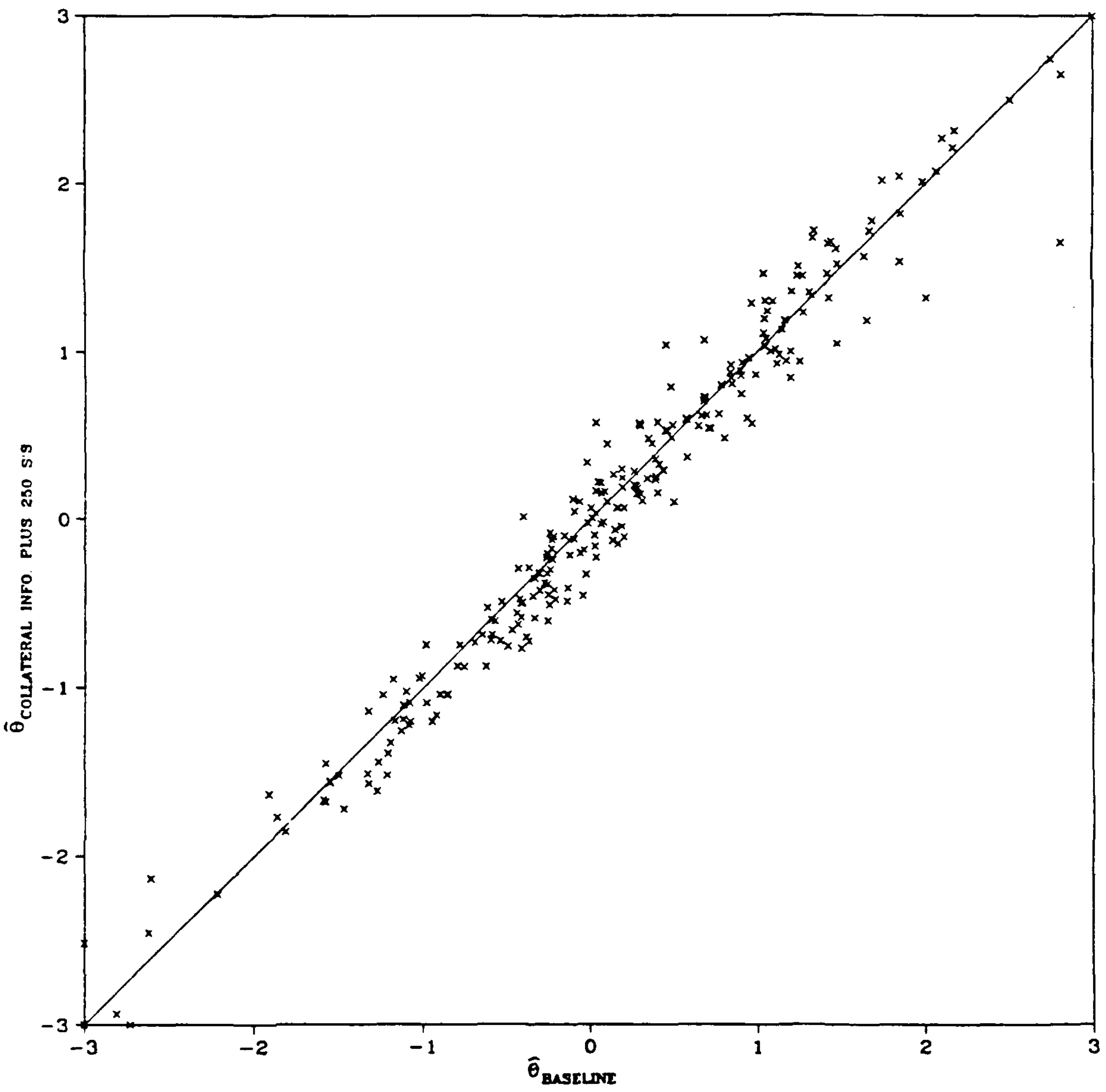


FIGURE 16

Examinee MLEs based on Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees

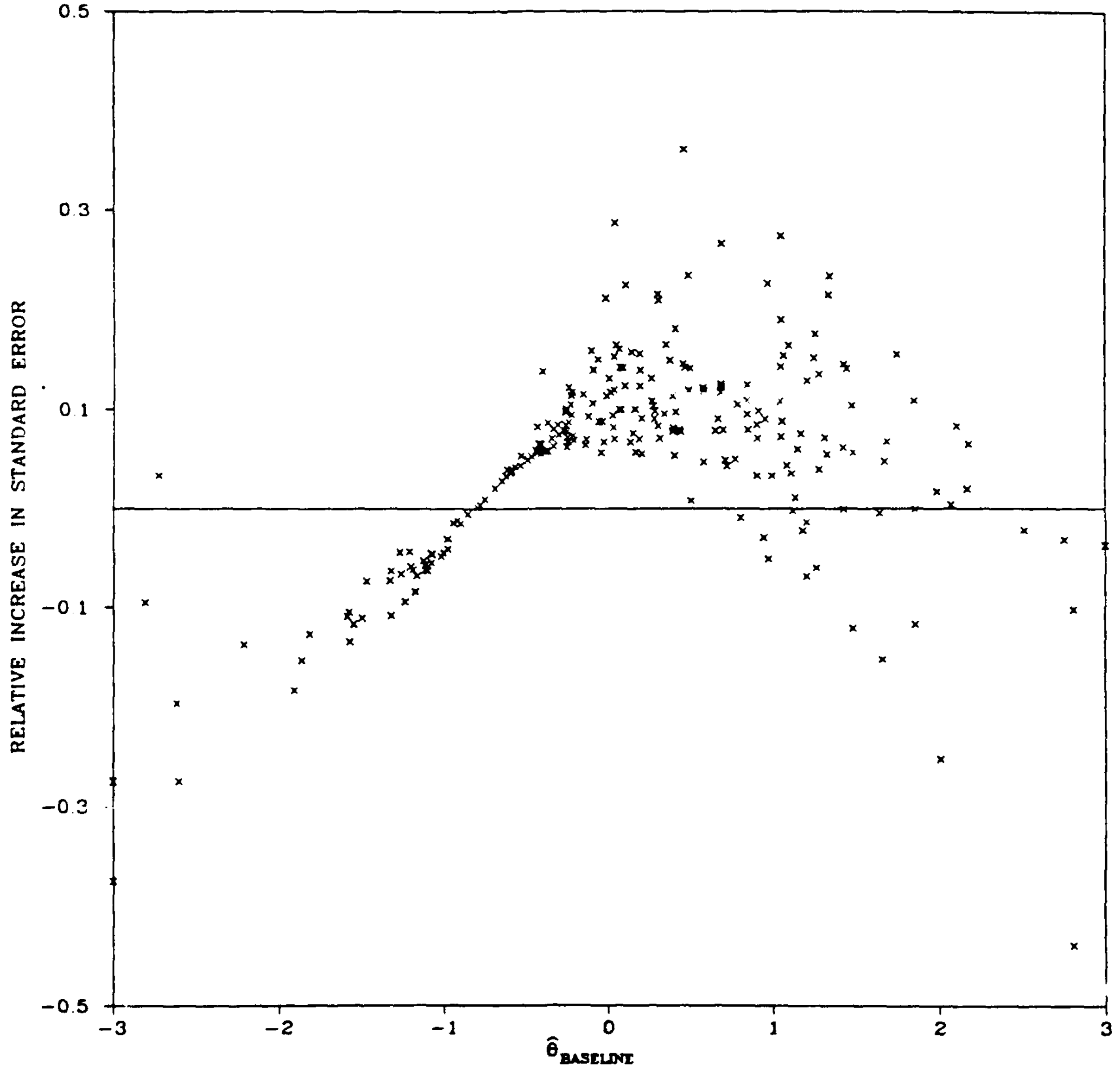


FIGURE 17

Comparison of Examinee Standard Errors Calculated with Baseline Estimates of Item Parameters and Expected Response Curves from Collateral Information and 250 Examinees

Dr. Terry Ackerman
Educational Psychology
260C Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Terry Allard
Code 1142CS
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217-5000

Dr. Nancy Allen
Educational Testing Service
Princeton, NJ 08541

Dr. Gregory Anrig
Educational Testing Service
Princeton, NJ 08541

Dr. Phipps Arabie
Graduate School of Management
Rutgers University
92 New Street
Newark, NJ 07102-1895

Dr. Isaac I. Bejar
Law School Admissions
Services
Box 40
Newtown, PA 18940-0040

Dr. William O. Berry
Director of Life and
Environmental Sciences
AFOSR/NL, NL Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Thomas G. Bever
Department of Psychology
University of Rochester
River Station
Rochester, NY 14627

Dr. Menucha Birenbaum
Educational Testing
Service
Princeton, NJ 08541

Dr. Bruce Bloxom
Defense Manpower Data Center
99 Pacific St.
Suite 155A
Monterey, CA 93943-3231

Dr. Gwyneth Boodoo
Educational Testing Service
Princeton, NJ 08541

Dr. Richard L. Branch
HQ, USMEPCOM/MEPCT
2500 Green Bay Road
North Chicago, IL 60064

Dr. Robert Brennan
American College Testing
Programs
P. O. Box 168
Iowa City, IA 52243

Dr. David V. Budescu
Department of Psychology
University of Haifa
Mount Carmel, Haifa 31999
ISRAEL

Dr. Gregory Candell
CTB/MacMillan/McGraw-Hill
2500 Garden Road
Monterey, CA 93940

Dr. Paul R. Chatelier
Perceptronics
1911 North Ft. Myer Dr.
Suite 800
Arlington, VA 22209

Dr. Susan Chipman
Cognitive Science Program
Office of Naval Research
800 North Quincy St.
Arlington, VA 22217-5000

Dr. Raymond E. Christal
UES LAMP Science Advisor
AL/HRMIL
Brooks AFB, TX 78235

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
Los Angeles, CA 90089-1061

Director
Life Sciences, Code 1142
Office of Naval Research
Arlington, VA 22217-5000

Commanding Officer
Naval Research Laboratory
Code 4827
Washington, DC 20375-5000

Dr. John M. Cornwell
Department of Psychology
I/O Psychology Program
Tulane University
New Orleans, LA 70118

Dr. William Crano
Department of Psychology
Texas A&M University
College Station, TX 77843

Dr. Linda Curran
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn, VA 22209

Dr. Timothy Davey
American College Testing Program
P.O. Box 168
Iowa City, IA 52243

Dr. Charles E. Davis
Educational Testing Service
Mail Stop 22-T
Princeton, NJ 08541

Dr. Ralph J. DeAyala
Measurement, Statistics,
and Evaluation
Benjamin Bldg., Rm. 1230F
University of Maryland
College Park, MD 20742

Dr. Sharon Derry
Florida State University
Department of Psychology
Tallahassee, FL 32306

Hei-Ki Dong
Bellcore
6 Corporate Pl.
RM: PYA-1K207
P.O. Box 1320
Piscataway, NJ 08855-1320

Dr. Neil Dorans
Educational Testing Service
Princeton, NJ 08541

Dr. Fritz Drasgow
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Defense Technical
Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
(2 Copies)

Dr. Richard Duran
Graduate School of Education
University of California
Santa Barbara, CA 93106

Dr. Susan Embretson
University of Kansas
Psychology Department
426 Fraser
Lawrence, KS 66045

Dr. George Engelhard, Jr.
Division of Educational Studies
Emory University
210 Fishburne Bldg.
Atlanta, GA 30322

ERIC Facility-Acquisitions
2440 Research Blvd., Suite 550
Rockville, MD 20850-3238

Dr. Marshall J. Farr
Farr-Sight Co.
2520 North Vernon Street
Arlington, VA 22207

Dr. Leonard Feldt
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard L. Ferguson
American College Testing
P.O. Box 168
Iowa City, IA 52243

Dr. Gerhard Fischer
Liebiggasse 5
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
U.S. Army Headquarters
DAPE-HR
The Pentagon
Washington, DC 20310-0300

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6800

Chair, Department of
Computer Science
George Mason University
Fairfax, VA 22030

Dr. Robert D. Gibbons
University of Illinois at Chicago
NPI 909A, M/C 913
912 South Wood Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Robert Glaser
Learning Research
& Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Susan R. Goldman
Peabody College, Box 45
Vanderbilt University
Nashville, TN 37203

Dr. Timothy Goldsmith
Department of Psychology
University of New Mexico
Albuquerque, NM 87131

Dr. Joseph McLachlan
Navy Personnel Research
and Development Center
Code 14
San Diego, CA 92152-6800

Alan Mead
c/o Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Timothy Miller
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

Dr. Ivo Molenaar
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Grote Kruisstraat 2/1
9712 TS Groningen
The NETHERLANDS

Dr. E. Muraki
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr Ratna Nandakumar
Educational Studies
Willard Hall, Room 213E
University of Delaware
Newark, DE 19716

Academic Progs. & Research Branch
Naval Technical Training Command
Code N-62
NAS Memphis (75)
Millington, TN 30854

Dr. W. Alan Nicewander
University of Oklahoma
Department of Psychology
Norman, OK 73071

Head, Personnel Systems Department
NPRDC (Code 12)
San Diego, CA 92152-6800

Director
Training Systems Department
NPRDC (Code 14)
San Diego, CA 92152-6800

Library, NPRDC
Code 041
San Diego, CA 92152-6800

Librarian
Naval Center for Applied Research
in Artificial Intelligence
Naval Research Laboratory
Code 5510
Washington, DC 20375-5000

Office of Naval Research,
Code 1142CS
800 N. Quincy Street
Arlington, VA 22217-5000
(6 Copies)

Special Assistant for Research
Management
Chief of Naval Personnel (PERS-01JT)
Department of the Navy
Washington, DC 20350-2000

Dr. Judith Orasanu
Mail Stop 239-1
NASA Ames Research Center
Moffett Field, CA 94035

Dr. Peter J. Pashley
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Circle, NW
Washington, DC 20036

Dept. of Administrative Sciences
Code 54
Naval Postgraduate School
Monterey, CA 93943-5026

Dr. Peter Pirolli
School of Education
University of California
Berkeley, CA 94720

Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

Mr. Steve Reise
Department of Psychology
University of California
Riverside, CA 92521

Mr. Louis Roussos
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Donald Rubin
Statistics Department
Science Center, Room 608
1 Oxford Street
Harvard University
Cambridge, MA 02138

Dr. Fumiko Samejima
Department of Psychology
University of Tennessee
310B Austin Peay Bldg.
Knoxville, TN 37966-0900

Dr. Mary Schratz
4100 Parkside
Carlsbad, CA 92008

Mr. Robert Semmes
N218 Elliott Hall
Department of Psychology
University of Minnesota
Minneapolis, MN 55455-0344

Dr. Valerie L. Shalin
Department of Industrial
Engineering
State University of New York
342 Lawrence D. Bell Hall
Buffalo, NY 14260

Mr. Richard J. Shavelson
Graduate School of Education
University of California
Santa Barbara, CA 93106

Ms. Kathleen Sheehan
Educational Testing Service
Princeton, NJ 08541

Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujisawa 251
JAPAN

Dr. Randall Shumaker
Naval Research Laboratory
Code 5500
4555 Overlook Avenue, S.W.
Washington, DC 20375-5000

Dr. Judy Spray
ACT
P.O. Box 168
Iowa City, IA 52243

Dr. Martha Stocking
Educational Testing Service
Princeton, NJ 08541

Dr. William Stout
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Kikumi Tatsuoka
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. David Thissen
Psychometric Laboratory
CB# 3270, Davie Hall
University of North Carolina
Chapel Hill, NC 27599-3270

Mr. Thomas J. Thomas
Federal Express Corporation
Human Resource Development
3035 Director Row, Suite 501
Memphis, TN 38131

Mr. Gary Thomasson
University of Illinois
Educational Psychology
Champaign, IL 61820

Dr. Howard Wainer
Educational Testing Service
Princeton, NJ 08541

Elizabeth Wald
Office of Naval Technology
Code 227
800 North Quincy Street
Arlington, VA 22217-5000

Dr. Michael T. Waller
University of
Wisconsin-Milwaukee
Educational Psychology Dept.
Box 413
Milwaukee, WI 53201

Dr. Ming-Mei Wang
Educational Testing Service
Mail Stop 03-T
Princeton, NJ 08541

Dr. Thomas A. Warm
FAA Academy
P.O. Box 25082
Oklahoma City, OK 73125

Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Douglas Wetzel
Code 15
Navy Personnel R&D Center
San Diego, CA 92152-6800

German Military
Representative
Personalstammamt
Koelner Str. 262
D-5000 Koeln 90
WEST GERMANY

Dr. Sherrie Gott
AFHRL/MOMJ
Brooks AFB, TX 78235-5601

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

Prof. Edward Haertel
School of Education
Stanford University
Stanford, CA 94305-3096

Dr. Ronald K. Hambleton
University of Massachusetts
Laboratory of Psychometric
and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

Dr. Patrick R. Harrison
Computer Science Department
U.S. Naval Academy
Annapolis, MD 21402-5002

Ms. Rebecca Hetter
Navy Personnel R&D Center
Code 13
San Diego, CA 92152-6800

Dr. Thomas M. Hirsch
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Paul W. Holland
Educational Testing Service, 21-T
Rosedale Road
Princeton, NJ 08541

Prof. Lutz F. Hornke
Institut fur Psychologie
RWTH Aachen
Jaegerstrasse 17/19
D-5100 Aachen
WEST GERMANY

Ms. Julia S. Hough
Cambridge University Press
40 West 20th Street
New York, NY 10011

Dr. William Howell
Chief Scientist
AFHRL/CA
Brooks AFB, TX 78235-5601

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Martin J. Ippel
Center for the Study of
Education and Instruction
Leiden University
P. O. Box 9555
2300 RB Leiden
THE NETHERLANDS

Dr. Robert Jannarone
Elec. and Computer Eng. Dept.
University of South Carolina
Columbia, SC 29208

Dr. Kumar Joag-dev
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright Street
Champaign, IL 61820

Professor Douglas H. Jones
Graduate School of Management
Rutgers, The State University
of New Jersey
Newark, NJ 07102

Dr. Brian Junker
Carnegie-Mellon University
Department of Statistics
Pittsburgh, PA 15213

Dr. Marcel Just
Carnegie-Mellon University
Department of Psychology
Schenley Park
Pittsburgh, PA 15213

Dr. J. L. Kaiwi
Code 442/JK
Naval Ocean Systems Center
San Diego, CA 92152-5000

Dr. Michael Kaplan
Office of Basic Research
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Jeremy Kilpatrick
Department of
Mathematics Education
105 Aderhold Hall
University of Georgia
Athens, GA 30602

Ms. Hae-Rim Kim
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Jwa-keun Kim
Department of Psychology
Middle Tennessee State
University
Murfreesboro, TN 37132

Dr. Sung-Hoon Kim
KEDI
92-6 Umyeon-Dong
Seocho-Gu
Seoul
SOUTH KOREA

Dr. G. Gage Kingsbury
Portland Public Schools
Research and Evaluation Department
501 North Dixon Street
P. O. Box 3107
Portland, OR 97209-3107

Dr. William Koch
Box 7246, Meas. and Eval. Ctr.
University of Texas-Austin
Austin, TX 78703

Dr. James Kraatz
Computer-based Education
Research Laboratory
University of Illinois
Urbana, IL 61801

Dr. Patrick Kyllonen
AFHRL/MOEL
Brooks AFB, TX 78235

Ms. Carolyn Laney
1515 Spencerville Rod
Spencerville, MD 20868

Richard Lanterman
Commandant (G-PWP)
US Coast Guard
2100 Second St., SW
Washington, DC 20593-0001

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
1310 South Sixth Street
University of IL at
Urbana-Champaign
Champaign, IL 61820-6990

Dr. Charles Lewis
Educational Testing Service
Princeton, NJ 08541-0001

Mr. Hsin-hung Li
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Library
Naval Training Systems Center
12350 Research Parkway
Orlando, FL 32826-3224

Dr. Marcia C. Linn
Graduate School
of Education, EMST
Tolman Hall
University of California
Berkeley, CA 94720

Dr. Robert L. Linn
Campus Box 249
University of Colorado
Boulder, CO 80309-0249

Logicon Inc. (Attn: Library)
Tactical and Training Systems
Division
P.O. Box 85158
San Diego, CA 92138-5158

Dr. Richard Luecht
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. George B. Macready
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Evans Mandes
George Mason University
4400 University Drive
Fairfax, VA 22030

Dr. Paul Mayberry
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. James R. McBride
HumRRO
6430 Elmhurst Drive
San Diego, CA 92120

Mr. Christopher McCusker
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Dr. Robert McKinley
Educational Testing Service
Princeton, NJ 08541

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 Winograd
Department of Psychology
Emory University
Atlanta, GA 30322

Dr. Martin F. Wiskoff
PERSEREC
99 Pacific St., Suite 4556
Monterey, CA 93940

Mr. John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Kentaro Yamamoto
03-0T
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Ms. Duanli Yan
Educational Testing Service
Princeton, NJ 08541

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

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