GRE®
RESEARCH

A General Bayesian Model for Testlets: Theory and Applications

Xiaohui Wang
Eric T. Bradlow
Howard Wainer

February 2002

GRE Board Professional Report No. 98-01P

ETS Research Report 02-02

ETS® Educational Testing Service
Princeton, NJ 08541

A General Bayesian Model for Testlets: Theory and Applications

Xiaohui Wang
University of North Carolina-Chapel Hill

Eric T. Bradlow
The Wharton School, University of Pennsylvania

Howard Wainer
Educational Testing Service

GRE Board Report No. 98-01P

February 2002

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board and Educational Testing Service.

Educational Testing Service, Princeton, NJ 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do not necessarily represent official Graduate Record Examinations Board position or policy.

****

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Educational Testing Service
Princeton, NJ 08541

Abstract

This paper extends earlier work (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000) on the modeling of testlet-based response data to include the situation in which a test is composed, partially or completely, of polytomously scored items and/or testlets. A modified version of commonly employed item response models, embedded within a fully Bayesian framework, is proposed, and inferences under the model are obtained using Markov chain Monte Carlo (MCMC) techniques. Its use is demonstrated within a designed series of simulations and by analyzing operational data from the North Carolina Test of Computer Skills and Educational Testing Service's Test of Spoken English (TSE). Our empirical findings suggest that the North Carolina Test of Computer Skills exhibits significant testlet effects, indicating significant dependence of item scores obtained from common stimuli, whereas the TSE exam does not.

Keywords: Bayesian hierarchical model, item response theory, polytomously scored items

Acknowledgements

The authors gratefully acknowledge the support of this research by the Graduate Record Examinations Board Research Committee, Educational Testing Service, and the College Board. We are also grateful to Yong-Won Lee and David Thissen, who provided us with advice, wisdom, and information functions for the summed scores of the North Carolina Test of Computer Skills, and to Catherine Hombo for a careful reading of an earlier draft of this manuscript. Lastly, we thank the Test of English as a Foreign Language (TOEFL) Program and the North Carolina Department of Public Instruction, which provided us with data from the Test of Spoken English and the North Carolina Test of Computer Skills, respectively.

Table of Contents

List of Tables

List of Figures

1 Introduction

Tests are typically made up of smaller components – items – which act in concert to measure the test designers' constructs of interest. Over the last half century one of the most important theoretical innovations in test theory has been a family of statistical models that characterize, in a stochastic way, the event of an examinee meeting an item (Birnbaum, 1968; Lord, 1952; 1980; Rasch, 1960). Underlying all versions of these item response models is the assumption that there exists an unobservable, latent proficiency (usually denoted θ) for each examinee which determines (in conjunction with parameters about the items) that examinee's likelihood of success on a given item. In practice, when utilizing item response models, it is always assumed that responses to all items are independent of one another after conditioning on the underlying, latent examinee proficiency. Experience has shown that when tests are made up of separate, unrelated items this assumption of conditional independence (CI) is sufficiently close to being true to allow these models to be of great practical usefulness. There are, however, some reasonably common circumstances in which the assumption of CI is not likely to be true.

The most frequent such circumstance observed in practice is when a test is constructed of "testlets." A testlet (Wainer & Kiely, 1987) is defined as an aggregation of items that are based on a single stimulus, such as in a reading comprehension test. In this case, a testlet might be defined as the passage and the set of four to 12 items that accompany the passage. It is not hard to imagine, therefore, that issues such as misinterpretation of the passage, subject matter expertise, fatigue, and so on would cause responses to these items to be more highly related than suggested by the overall (omnibus) latent proficiency for the entire test. In some sense, this lack of CI is a form of unidimensional proficiency model misfit, which may be explainable by the test structure (i.e., the testlet design). It is the incorporation of this test design structure into a formal probability model that motivated this research.

Much previous work exists in the psychometric literature on the modeling and/or detection of testlet dependence. Under the heading of "appropriateness measurement," Drasgow, Levine, and Williams (1985) and Levine and Drasgow (1988) describe parametric approaches for identification of deviations from standard unidimensional item response models. In more recent work, Stout (1987) and Zhang and Stout (1999), develop nonparametric approaches (e.g., the DETECT statistic) to determine when proficiency unidimensionality is likely to be violated. The previous research most relevent to the present work–by Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow, and Du (2000)–proposes a parametric Bayesian model for item test scores composed of a mixture of *binary* independent and testlet items. Their base models, which are a modification of standand item response models (two and three parameter logistic models) that include an additional interaction term for persons answering a given testlet, demonstrate that (a) both examinee proficiencies and item parameters are biased when testlet dependence is ignored; (b) the amount of testlet dependence varies across testlets; and (c) testlet dependence exists in operational tests. However, their work assumes that each item response is binary; in the current study, we address an extension to that assumption motivated by the increasing number of operational tests that are composed of a mixture of binary and polytomous items, both independent and nested within testlets.

This extension is important for practical as well as theoretical reasons. First, tests with this format are in use, and existing scoring models that do not take into account within-testlet dependence will yield overly optimistic estimates of the test's overall precision. Therefore, we expect that mixed format tests with testlets models (a term we coin for tests composed of a mixture of binary and polytomous items, some independent and some within testlets) and related estimation procedures can have immediate operational use, such as with Educational Testing Service's (ETS) Test of Spoken English (TSE) and various achievement exams given very widely.

Secondly, some recent research has shown that richer inferences and diagnostic proficiency in-

formation can be obtained using portfolios (Advanced Placement Studio Art), essays, and other types of constructed responses, which are scored polytomously. As this belief becomes more accepted, and as such test items become more common, we expect models with the capability of the one proposed here to be widely used. In addition, as testlets composed of one item are single dependent items, by definition, our model easily simplifies to the standard model assuming CI.

The remainder of this manuscript is laid out as follows. In Section 2, we describe in detail our Bayesian parametric model for mixed binary-polytomous test with testlets. Section 3 contains a description of the computational approach utilizing a Markov chain Monte Carlo (MCMC) sampler. A large scale simulation study demonstrating the efficacy of our approach under a wide range of realistic test conditions is provided in Section 4. In Section 5 we apply our model to operational data from the North Carolina Test of Computer Skills (details can be found in pdf "Assessment brief" at http://www.dpi.state.nc.us/accountability/testing/computerskills/index.html) and ETS's Test of Spoken English (TSE) exam (Educational Testing Service, 1995). These applications demonstrate both the existence of testlet effects in some cases and not in others, and the operational feasibility of our approach. Summary conclusions are given in Section 6. A small technical appendix with details of our implementation of an MCMC sampler is also provided.

2 The Model

As our model must encompass both binary and polytomous items, two basic (and widely used) probability kernels drove our approach: the three-parameter logistic model for binary items (Birnbaum, 1968), and the polytomous item response model introduced by Samejima (1969). These models are given respectively by:

$$p_{ij}(1) = P(y_{ij} = 1|\theta, \omega_j) = c_j + (1 - c_j)\text{logit}^{-1}(t_{ij}), \quad \text{and} \tag{1}$$

$$p_{ij}(r) \quad = \quad P(y_{ij} = r|\theta, \omega_j, d) = \Phi(g_r - t_{ij}) - \Phi(g_{r-1} - t_{ij}) \tag{2}$$

where $p_{ij}(r)$ denotes the probability that examinee $i = 1, \ldots, I$ receives score $r = 1, \ldots, R_j$ on item $j = 1, \ldots, J$ (e.g., $p_{ij}(1)$ is the probability of a correct binary item); c_j is the lower asymptote ("guessing" parameter) for binary item j; ω_j is the set of item j parameters; g_r is the latent cutoff for the polytomous items such that observed score $y_{ij} = r$ if latent score $s_{ij} = t_{ij} + \epsilon_{ij}$ satisfies $g_{r-1} < s_{ij} \leq g_r$ (the set of cutoffs is denoted g); ϵ_{ij} is a standard unit Gaussian random variable; Φ is the normal cumulative density function; $logit(x) = log(x/(1-x))$; and t_{ij} (described below) is the latent linear predictor of score [1].

We used the Birnbaum model for the binary items because it encompasses simpler two-parameter and Rasch models as special cases, and because it accounts for the probability that examinees of very low ability may still answer a binary item correctly due only to chance. Since we incorporated the possibility of "turning off" the extra features of the Birnbaum model in our program, and the data can inform about the level of complexity of model needed, utilizing a more general structure seemed warranted. The Samejima model for polytomous items has a nice intuitive explanation as a latent true score model for examinee-item combination ij. That is, when examinee i is confronted with item j, that examinee responds with latent proficiency centered around his or her true score t_{ij}

[1] We use the term "score" here in an imprecise but comfortable way. It would be more accurate to say "examinee i receives a rating of r on item j", since typically polytomous models are used to score items that have been rated. But not always. Sometimes (p. 17) a testlet is merely a set of dichotomous items and the value of r is the number of right answers within the testlet. In addition, we call the expected value of the latent distribution the "true score" although in traditional true score theory it is the expected value of the observed score distribution that is called the true score. We have taken these imprecise liberties because trying to maintain this verbal precision led to very arcane descriptions and most importantly, because our mathematics says exactly what we mean. For a more complete description of test scoring that uses terminology more carefully, see Thissen & Wainer (2001).

with random error ϵ_{ij}. Observed score $y_{ij} = r$ occurs when the latent score is within an estimated (latent) range $[g_{r-1}, g_r]$. The ability of our approach to model extra dependence due to testlets, first described in Bradlow et al. (1999), is obtained by extending linear score predictor t_{ij} from its standard form:

$$t_{ij} = a_j(\theta_i - b_j) \tag{3}$$

where a_j, b_j, and θ_i have their usual interpretations as item slope, item difficulty, and examinee proficiency to:

$$t_{ij} = a_j(\theta_i - b_j - \gamma_{id(j)}) \tag{4}$$

where $\gamma_{id(j)}$ is the testlet effect of person i to item j nested in testlet $d(j)$. The extra dependence of items within the same testlet for a given examinee is modeled in this manner, as both would share the effect $\gamma_{id(j)}$ in their score predictor. By definition, $\gamma_{id(j)} = 0$ for all independent items, or testlets of size one.

Using these kernels as a base, we supposed the following general testing set-up to fully explicate our model. Suppose I examinees each take an examination composed of J items, where $J = J_b + J_p$, and J_b is the number of binary items in the test while J_p is the number of polytomous items. Furthermore, let $\mathcal{J}_b$ denote the set of binary items and $\mathcal{J}_p$ the set of polytomous items. We further suppose that each of the J items are nested within K testlets; that is, $d(j) \in \{1, \ldots, K\}$, where $k_{d(j)}$ and $\mathcal{K}_{d(j)}$ denote the number of items and sets of items nested within testlet $d(j)$. Under this paradigm, and using the probability kernels given in equation (1) and (2), we obtain the likelihood for observed test score matrix $\mathbf{Y} = (y_{ij})$ given by:

$$P(Y|\Lambda_1) = \prod_{i=1}^{I}\left\{\prod_{j\in\mathcal{J}_b}\left[\frac{\frac{e^{q_j}}{1+e^{q_j}}+e^{a_j(\theta_i-b_j-\gamma_{id(j)})}}{1+e^{a_j(\theta_i-b_j-\gamma_{id(j)})}}\right]^{y_{ij}}\left[\frac{1-\frac{e^{q_j}}{1+e^{q_j}}}{1+e^{a_j(\theta_i-b_j-\gamma_{id(j)})}}\right]^{1-y_{ij}}\right. \tag{5}$$
$$\left.\cdot\prod_{j\in\mathcal{J}_p}\prod_{r=1}^{R_j}\left(\Phi(g_r-t_{ij})-\Phi(g_{r-1}-t_{ij})\right)^{1(y_{ij}=r)}\right\}$$

where $q_j = \text{logit}(c_j)$; $\Lambda_1 = \{\vec{\theta}, \vec{a}, \vec{b}, \vec{q}, \vec{g}, \vec{\gamma}\}$, the set of likelihood parameters; and $1_{(.)}$ is an indicator function. The quessing parameter c_j is transformed to the logit scale as we assert a Gaussian prior (given below) for its effect. We note that the likelihood given in (5) does assume CI across persons and items but only after conditioning on overall examinee proficiency θ_i *and* testlet effect ("proficiency") $\gamma_{id(j)}$, a much less restrictive assumption.

As in Bradlow et al. (1999), we embedded the model described in equation (5) in a larger Bayesian hierarchical framework. This framework allows for borrowing of information across examinees, items, and most importantly (to this research), testlets in a setting in which a large degree of commonality is likely to exist (Gelman, Carlin, Stern, & Rubin, 1995). In addition, it allowed us to properly model the uncertainty in these quantities. We expected this Bayesian framework to add substantially to the mixed testlet model, for while it is true that each examinee typically answers many items (or at least enough by design to pin down his or her ability), and each item is answered by many examinees, each person-testlet combination has sparse information and will benefit greatly from the Bayesian paradigm.

We asserted the following prior distributions for Λ_1:

$$\begin{aligned} a_j &\sim N(\mu_a, \sigma_a^2) \\ b_j &\sim N(\mu_b, \sigma_b^2) \\ q_j &\sim N(\mu_q, \sigma_q^2) \end{aligned} \tag{6}$$

$$\theta_i \sim N(0,1)$$

$$\gamma_{id(j)} \sim N(0, \sigma^2_{d(j)})$$

where $N(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. We note that the mean and variance of the ability distribution are fixed at 0 and 1 (as is standard) to identify the model, and that the variance of the testlet effects, $\sigma^2_{d(j)}$, is testlet specific, allowing the amount of excess dependence across testlets to vary. We denoted the set of parameters for the priors by $\Lambda_2 = \{\mu_a, \mu_b, \mu_q, \sigma^2_a, \sigma^2_b, \sigma^2_q, \sigma^2_{d(j)}\}$ and the full set of model parameters by $\Lambda = \Lambda_1 \bigcup \Lambda_2$ with elements λ_k. We also let Λ_{-k} denote the set of all elements of Λ excluding the k-th element.

To complete our model specification, we added a set of hyperpriors for parameters Λ_2 given in (6) to reflect the uncertainty in their values. The distributions for these parameters were chosen out of convenience as conjugate priors to Λ_1. For the distribution means we selected $\mu_a \sim N(0, V_a)$, $\mu_b \sim N(0, V_b)$, and $\mu_q \sim N(0, V_q)$, where $V_a^{-1} = V_b^{-1} = V_q^{-1}$ were set to 0. Slightly informative hyperpriors (to ensure proper posteriors) were used for all prior variances given by $\sigma^2_z \sim \chi^{-2}_{g_z}$, an inverse chi-square random variable with g_z degrees of freedom, where $g_z = 0.5$ for all distributions. It is well established (Albert & Chib, 1993; Bradlow & Zaslavsky, 1999) that the marginal posterior distributions for the elements of Λ, $p(\lambda_k|Y)$, are not obtainable in closed-form. That is, being representable as the product of the mixed testlet likelihood given in equation (5), the priors given in (6), and the hyperpriors are not integrable analytically.

3 Computation

To facilitate computation for this model, we implemented an MCMC computational approach. To draw inferences from the marginal posteriors $p(\lambda_k|Y)$, we obtained samples from their distributions using an MCMC sampler (Gelfand & Smith, 1990; Roberts & Smith, 1993). Inferences based on

posterior means, quantiles, interesting posterior probabilities, and so on were then derived from sample-based estimates. The standard MCMC approach–to begin with some starting value, $\Lambda^{(0)}$, and then iterate by sampling in turn from the set of conditional distributions:

$$\begin{aligned} & p(\lambda_1^{(t+1)}|Y, \Lambda_{-1}^{(t)}) \\ & p(\lambda_2^{(t+1)}|Y, \lambda_1^{(t+1)}, \Lambda_{-1,-2}^{(t)}) \\ & \vdots \\ & p(\lambda_K|Y, \Lambda_{-K}^{(t+1)}) \end{aligned} \tag{7}$$

for $t = 0, \ldots, M$ iterations until convergence, and then some desired number thereafter was non-trivial in this case due to the inability to sample directly from the conditional distributions corresponding to the likelihood parameters Λ_1. The conditional distributions corresponding to Λ_2 could be sampled directly, as they were chosen (as mentioned earlier) by convenience to be conjugate to Λ_1. To sample from the subcomponents of Λ_1 we applied two different approaches depending on the parameter of interest. That is, we utilized one sampling approach for the latent polytomous cutoffs g_r, and one for the remaining parameters in Λ_1. Complete details of these approaches are provided in the Appendix; we briefly describe them here.

For the latent cutoffs g_r corresponding to the polytomous items, rather than sample from the conditional distributions $p(g_r|Y, \Lambda_{-g_r})$ as in (7), we instead utilized a different set of conditional distributions which augment parameter vector Λ with latent linear score matrix $S = (s_{ij}) = t_{ij} + \epsilon_{ij}$ as given in equation (4). The advantage of this data augmentation approach (Tanner & Wong, 1987) is that, while the distributions for $p(g_r|Y, \Lambda_{-g_r})$ can't be sampled directly, $p(g_r|Y, \Lambda_{-g_r}, S)$ can. Since the data augmentation approach is not directly available for the conditional distributions involving the remaining subcomponents of Λ_1, we sampled from their conditional distributions using a Metropolis-Hastings step (Hastings, 1970) with normal sampling densities. The means of the

sampling densities were set to the previously drawn value $\lambda_k^{(t)}$, and the variance was set adaptively to achieve a high acceptance rate (Gelman, Roberts, & Gilks, 1996). Although, a single approach (Metropolis-Hastings) could have been used to sample all subcomponents of Λ_1, a comparison (not reported) between Metropolis-Hastings for all parameters and the data augmentation approach for the cutoffs indicated that applying both algorithms would be superior.

4 Model Testing

To check both accuracy and computational time for realistic size data sets, and to understand the impact of dependence in the mixed testlet model, we conducted extensive simulations using our computational approach. We performed our simulation study for two primary purposes. First, we wanted to confirm that we could obtain accurate estimates of the model parameters under wide variation of three experimental factors; that is, we sought realistic values likely to be seen in practice. As this procedure will have operational utility, practical application of this methodology was of critical importance. Second, we wished to compare our results with current standard estimation approaches in the testing industry specifically, those using MULTILOG (Thissen, 1991).

In testing the model and our estimation approach, we considered three experimental factors, denoted Factors A, B, and C. Factor A was the number of categories (n_c) for each item. Thus for dichotomous items $n_c = 2$, and for polytomous items, we manipulated the number of score levels $n_c > 2$. Factor B, testlet length or n_t, was defined as the number of items within a testlet. Finally, factor C, the variance of the testlet effects, $\sigma^2_{d(j)}$, indicated the degree of within-testlet dependence, as in (6).

Nine different simulation conditions were studied, using a data generation computer program developed for this purpose. For each of the nine conditions, five data sets were simulated indepen-

dently for a total of 45 simulated data sets. Each data set consisted of responses of 1,000 simulees to a test of 30 items. Among those 30 items, 12 were independent dichotomous items (i.e., not within testlets), and the remaining 18 were dichotomous and polytomous testlet items. Of course, as the 12 independent, binary items contained no testlet effect by definition ($\sigma^2_{d(j)} = 0$), and had a fixed number of categories ($n_c = 2$), our simulation manipulations corresponded to variations in the remaining 18 items. This test design was utilized to mimic many current operational tests in which independent binary multiple-choice items are followed by essay/portfolio testlets (e.g., Advanced Placement Program tests).

Because a 3^3 design with 27 combinations would have been too computationally "expensive" to run, and because our interests predominantly lay in estimating main effects and two-way interactions, a Latin Square design was employed to cover the variation of these three factors. For Factor A, we chose the number of response categories to be 2, 5, or 10, corresponding to binary items and to polytomous items scored on 5 and 10 point scales, respectively. Thus, if $n_c = 2$, the entire test is binary (including the 12 independent items and the 18 testlet items), and this mimics the work of Wainer, Bradlow, and Du (2000). For Factor B, we designated the length of the testlets as three items, six items, and nine items, respectively. Since we fixed the total number of testlet items at 18, these assignments corresponded to six testlets, three testlets, and two testlets. Therefore, Factor A times Factor B yielded nine combinations. For Factor C, the variance of testlet effects, we chose its values to 0.0 (i.e., no testlet effect), 0.5 (small variance), and 1.0 (bigger variance). Since $\theta \sim N(0, 1)$ to identify the model, all values of $\sigma^2_{d(j)}$ are relative to 1, the variance of the person abilities. Using the Latin Square design, we let Factor C change for each level of Factors A and B. Table 1 describes our simulation design.

From previous work of Wainer, Bradlow, and Du (2000), we knew that this model (and the MCMC estimation approach) containing only binary items (conditions 1, 4, and 7) works very well

Factor C		Factor A		
Variance of Testlet Effect		2 categories	5 categories	10 categories
	6 testlets	0.0 (1)	0.5 (2)	1.0 (3)
Factor B	3 testlets	0.5 (4)	1.0 (5)	0.0 (6)
	2 testlets	1.0 (7)	0.0 (8)	0.5 (9)

Note: Numbers from 1-9 in parentheses indicate our labeling of the simulation conditions and are used throughout.

Table 1: Table of Simulation Design

for situations with a positive and a zero testlet effect. Under the design in Table 1, we expected that our procedure would give accurate estimates for the item parameters, including the cutoffs g_r for the categories of polytomous items, regardless of the number of categories n_c, the testlet length n_t, and the variance of the testlet effects.

To make the simulated test data as similar as possible to real world applications, we selected population distributions for the parameters used to generate the data that correspond to previous analyses of the SAT examination. Specifically, we used $\theta_i \sim N(0,1)$, $a_j \sim N(1.5, 0.45^2)$, $b_j \sim N(0,1)$, and $c_j \sim N(0.14, 0.05^2)$. The population distributions for a_j were left-truncated at 0.3, and those for c_j were left-truncated at 0.0 and right-truncated at 0.6. [2]

[2]For practical usage, when a_j is too small (as estimated from calibration samples), that implies a very low item discrimination (i.e., the item slope is too low, indicating that the item is unable to meaningfully differentiate people with varying abilities) and such items are never used; they are pruned from the test. Similarly, as c_j are guessing parameters, when items are guessed correctly too often, the items are pruned. Such truncations were not critical to our simulation design and did not occur that frequently; equally accurate results were obtained in test runs without these restrictions.

All 45 simulated data sets were analyzed by our procedure. The estimated parameters were then compared in various ways with the true model parameters to determine whether the procedure recovered their values. The estimated parameters were posterior means of the MCMC draws for each parameter obtained from the last 1,000 draws of a single MCMC chain of length 3,000. The initial 2,000 draws were discarded as an initial burn-in period.

We assessed the performance of our model using two criteria: the correlation between the estimated parameters and the true parameters, and the mean square error of the estimates from the true values. The full results are presented in Table 2 and Table 3, respectively. Each value in these tables represents an average over the five replications for that condition. For ease of presentation, the values in Table 3 are multiplied by 100. Appended to Table 2 are rearranged subtables (Tables 4, 5 and 6), which explicitly reflect the main effects embedded in the structure of the experimental design. We have only included those tables in which the main effects varied meaningfully with the values of an independent variable.

INSERT TABLES 2, 3, 4, 5, 6 HERE

Table 2, shows that our procedure provided very accurate estimates for item difficulty (b) and cutoff (g_r) parameters (average correlations of 0.992 and 0.980 respectively) across all simulation conditions. The average correlation equals 0.93 for the ability parameters θ and 0.89 for the discrimination (a) parameters. As is typical with IRT models, the procedure's ability to estimate the guessing parameters was more modest; in this case, the average correlation is 0.60. This was expected because there are very few simulees whose proficiency is low enough to provide information in the estimation of the guessing parameter (c). The magnitude of the correlations were all consistent with our prior beliefs based on past research. The differences in these correlations across the various levels of the design were small, but consistent, for some parameters.

In Table 4, we see that the accuracy of estimation of the slope parameter (a) increased with the number of items within a testlet (since test length was fixed, as the number of testlets n_t was reduced their average length was, perforce, increased). This was expected since with longer testlets there is more data available for the slope estimates. We also found increased precision as the number of categories, n_c, for each polytomous item increased.

We found a similar effect on the difficulty parameter (b), shown in Table 5. This parameter was so well estimated that it was difficult for any variation in the independent variables to have much effect. Last, when we examine the effect of varying the testlet parameter–var (γ) in Table 6–we find that only slope (a) and proficiency (θ) showed any consistent effect (albeit not statistically different); in both cases, increasing testlet effect decreased precision.

An analysis of the mean square errors (Table 3) shows almost exactly the same pattern. Using our model, $\hat{t}$ item parameters were well estimated, with average squared prediction error ranging from a low of 0.006 (0.60 x 10^{-2} from Table 3) for the difficulty parameters, to a high of 0.054 (5.4 x 10^{-2} from Table 3) for the guessing parameters. Simulee abilities were also extremely well estimated, with MSE equal to 0.001 (1.0 x 10^{-2}). The same main effects between precision and independent variables shown in Table 2 and its subtables reappeared in Table 3, but now when considering mean square error (MSE).

5 Two tests in need of a scoring model

Next, we applied our approach, now validated, to the analysis of operational data from two tests made up of testlets: the North Carolina Test of Computer Skills, one section of which is composed of four testlets that turned out to show very large testlet effects, and the TSE exam, which is composed of four testlets that manifested essentially none of the excess local dependence that is typical of

testlet-based tests. Our analyses demonstrate how the use of our new model allows accurate scoring when its generality is needed, and how it still provides important and useful information even when its generality appears not to be needed.

Test 1. North Carolina Test of Computer Skills. The North Carolina Test of Computer Skills is given to 8th graders and must be passed as a requirement for graduation from junior high school. It was developed to help ensure basic computer proficiency for graduates of the North Carolina Public Schools, and is made up of two parts. The first part of the exam is presented in a standard multiple-choice format, and the second part is performance based, consisting of four testlets that deal with keyboarding, word processing/editing, database use, and spreadsheet use. The keyboarding portion includes three polytomous items scored on a four-point scale, while the word processing/editing, database, and spreadsheet testlets include six to 10 items scored either dichotomously or trichotomously. Each student receives two separate scores–one for the multiple-choice portion and one for the performance portion–and is required to pass both parts.

In their analysis of the performance section, Rosa, Swygert, Nelson, and Thissen (2001) found that that the reliability of the computer skills test, assuming no testlet effects, was 0.83, whereas when the test was scored as being made up of four testlets, its reliability was 0.65. If all of the items measure the same trait and there is no excess local dependence (testlet effect), we would expect that an estimate of the test's reliability based strictly on the items (ignoring the testlet structure) would be the same as an estimate based on the testlets. The result obtained suggests that there is substantial within-testlet dependence. To assure an honest estimate of the precision of the test, it would seem that some other test scoring model, beyond a standard model assuming CI, should be used.

Our testlet model is sufficiently general to allow the entire exam to be scored together. Such an approach has some important advantages when a total exam, like this one, is predominantly

unidimensional (Rosa et al. 2001). Principal among these advantages is that a single score for the two parts combined would yield a more reliable measure of a student's computer proficiency.

Although providing a single score was technically feasible, the strategy of computing two separate scores and requiring the student to pass both parts was adopted for at least two reasons–the first economic, the second technical. The economic reason was that the performance portion of the test is very expensive to administer, so the students take the cheap part (the multiple-choice section) over and over again until they pass it. Then they take the performance part until they pass it. The hope was that the extra study involved in the retesting of the multiple-choice section would reduce the number of times that the performance part needed to be taken. There is some evidence that this is true. At the very least, those that never pass the first part never take the second. The technical reason for providing two scores was that, until now, no rigorous scoring model was available that could mix all parts together in an optimal fashion–although there certainly were methods for doing it fairly well (Rosa et al. 2001).

We fit the testlet model given in Equation (5) and (6) to the performance data from one administration of the North Carolina test. The test was administered during the 1994-95 school year as part of an item tryout. The tryout included potential performance items arranged into 12 forms numbered 13-24; the data used in the current study was from form number 13, which included 26 items divided into testlets as shown in Table 7. The sample size for this form was 266, roughly one twelfth of the 3,099 examinees with complete data in the field test.

An MCMC sampler was run from three starting points for 3,000 draws each. The first 2,000 draws from each chain were discarded, and the remaining draws were used for inference. The last column of Table 7 presents the estimated values of the testlet effects var(γ). The interpretation of the size of the testlet effects is aided by remembering that they are on the same scale as examinee proficiency. Thus a testlet effect of one means that the variance associated with local dependence

is of the same order of magnitude as the variance of examinees' ability. We see that there were very large testlet effects for the word processing/editing portion of the test as well as the spreadsheet section. This reflects the highly interconnected nature of these tasks. There was a smaller, but nevertheless substantial, effect for database use. Only the keyboarding section seemed to yield independent items.

INSERT TABLE 7 HERE

The simulation results discussed in Section 4 indicated that having a substantial testlet effect will not affect the accuracy of some of the parameters, but it will affect testlet and total test information; $I(\theta) = -E(\partial^2 \log L / \partial \theta^2)$, where L is (for our model) the mixed testlet model likelihood given in Equation (5). It is worthwhile to compare the results for test information we obtained from our model with what was yielded by two traditional approaches. As expected, a comparison of estimated proficiencies $\hat{\theta}$ across methods indicated a correlation above 0.95 and hence are not reported in detail here. Test information in test theory, as in many applications, is critical, in that it informs the level of certainty of ability estimation at varying levels of ability.

Figure 1 presents three information curves. The highest curve was obtained by fitting an IRT model to the individual items of the North Carolina test and assumed that the items are conditionally independent (setting var(γ) equal to zero), and by using MULTILOG. The middle curve was estimated using our MCMC output. Specifically, $I(\theta)$ was computed pointwise for 100 equally spaced grid points of θ between -3 and +3 for each of the draws of the sampler. The value for $I(\theta)$ under our model for each grid point was then computed as average information over the draws. The computational formula for $I(\theta)$ for each item under our mixed-testlet model was easily derived from Equation (5), and can be shown to be equal to

$$I_j(\theta) = \sum_{r=1}^{R_j} \frac{(p_{ij}^{*\prime}(r-1) - p_{ij}^{*\prime}(r))^2}{p_{ij}^{*}(r-1) - p_{ij}^{*}(r)} \quad (8)$$

a special case of the formula given in Baker (1992; page 241, equation 8.19), where $p_{ij}^{*}(r) = \sum_{r'>r} p_{ij}(r)$ with $p_{ij}(r)$, as given in Equation (2) for polytomous items and in Equation (1) for dichotomous ones, and $p_{ij}^{*\prime}(r)$ is its respective derivative. That is, $p_{ij}^{*}(r)$ is the cumulative probability of being greater than r under the model, and hence, $p_{ij}^{*}(r) - p_{ij}^{*}(r-1) = p_{ij}(r)$. The lowest curve was obtained by treating each testlet as a single polytomous item and only recording the total number of points assigned, also using MULTILOG. This latter approach has been widely used (Thissen, Steinberg, & Mooney, 1989; Wainer, 1995), but as we see here, it tends to be too conservative; by using only the total score it loses any information that is carried in the exact pattern of responses. In summary, Figure 1 indicates that ignoring the testlet effect provides standard errors inversely related to test information that will be potentially too small under the CI assumption and too large when collapsing testlet data into a single score.

INSERT FIGURE 1 HERE

One further finding from the North Carolina analysis bears mention. Figure 2 shows information curves for each of the testlets, as well as the total test's information curve. This display makes absolutely clear the relationship between testlet topic and the proficiency levels at which that topic provides information. The word processing/editing section of the North Carolina test provides its peak information for examinees at the lowest proficiency levels, whereas the section on database use is focused at the highest proficiency levels. Interestingly, the limited value of the keyboarding section is distributed pretty uniformly across the entire proficiency range. These findings of highly differentiated testlets are in stark contrast to our findings for the Test of Spoken English, shown next.

INSERT FIGURE 2 HERE

Test 2. The Educational Testing Service's Test of Spoken English. A second example of the same sort of testlet design manifests itself in the TSE exam, the primary purpose of which is to measure the ability of nonnative English speakers to communicate orally in English. TSE scores are widely used by North American institutions for the selection of teaching assistants and doctoral students. They are also used outside of academia in many selection and certification circumstances, most commonly in the health professions for physicians, nurses, pharmacists, and veterinarians.

The TSE is made up of some independent items and three testlets, which themselves are composed of polytomous items[3]. Each testlet requires a particular language function (i.e., narrating, recommending, persuading), and is composed of a stimulus (e.g., a map, a sequence of pictures, a graph) and related items. After having the opportunity to study the stimulus, a series of orally presented questions about that stimulus are posed. The test is delivered on an audio tape augmented by a test booklet, and the examinee's responses are recorded on a separate answer tape.

Each TSE item is scored by two expert raters on a nine-point rating scale. If raters differ by more than one point on average over the 12 items scored, a third rater is brought in to adjudicate. The ratings of the two closest raters are then averaged and summed to provide the final score.

We fit our model to Form 3VTSO1 of the TSE, which was administered in January of 1999. A total of 2,127 individuals took the test at that time. We transformed their scores onto a scale that ranges from 20 to 60, with the various levels interpreted in terms of the ability to effectively

[3]Of the 12 items on the test: Testlet I is made up of items 1 through 4 which are based on the same map; Testlet II is made up of items 5 through 8, which are based on a series of pictures; item 9 is a discrete item that asks the examinee to summarize some information of the speaker's own choosing; Testlet III is made up of items 10 and 11, which are based on the same graph; and item 12 is a discrete item that provides a train schedule and requires examinees to give instructions to someone who needs to get somewhere.

communicate in English. These levels are shown

Score Level	Communication Ability In English
60	Almost always effective
50	Generally effective
40	Somewhat effective
30	Not generally effective
20	No effective communication

Our results indicate that the current practice of having raters score each item and then just adding them up as if they were independent is not unreasonable. We reached this conclusion when we found that there was essentially no excess local dependence on this form of the TSE [$\text{var}(\gamma) < .04$ for all testlets]. Because the size of the testlet effects was so small, we concluded that the current practice of ignoring it when calculating test summaries was completely justified for this form of the test. We can only speculate about whether all TSE forms show this same characteristic. Our experience with other tests (e.g., the North Carolina Test of Computer Skills and the Law School Admissions Test, to pick two) suggests that an absence of testlet effects is the exception, not the rule. In a separate research project currently underway, we are collecting testlet covariates (e.g., passage length, topic domain, and so on) to aid test developers in creating priori assessments of which testlets are likely to violate of CI (and their extent). As ultimately total test information at a minimum level is desired, this should have great practical importance.

After fitting our model to the TSE data, we used MCMC draws to construct the expected score curves for each item, $E(y_{ij}|\Lambda_1^{(t)})$. Figure 3 shows the expected score curves (averaged over the draws $\Lambda_1^{(t)}$) for three items: the easiest (item 1), the hardest (item 9), and the item of median difficulty (item 5). Each of these items was meant to test different aspects of English proficiency, which were

anticipated to become increasingly sophisticated as the test progressed. As Figure 3 shows, gigantic differences in expected score did not occur at any level of proficiency, but the biggest differences occured at a very low level of proficiency ($\theta = -1.75$). This could be seen more clearly when we plotted the difference in expected score between item 1 and item 9 in Figure 4. As is expected, at very low and very high levels of proficiency there are no differences in performance among the items. At the prior mean proficiency level ($\theta = 0$), there is only a four point difference in expected score between the most difficult and easiest item.

INSERT FIGURES 3 and 4 HERE

It turns out that all other items fall within this envelope. One interpretation of this result is that, once an individual's proficiency reaches a level characterized as "somewhat effective" the various aspects of linguistic proficiency spanned by this test are of almost equal difficulty. This was apparently suspected by the language experts who construct the TSE exam, but they had never been able to find compelling evidence to support this suspicion. Our model provided this evidence.

As part of our analysis, we also looked at the information function, as in Equation (8), for the entire TSE test. We found that the area of peak accuracy of the test is remarkably broad (see Figure 5). Thus, this form of the TSE yields equally accurate measurement across a very broad spectrum of examinee proficiencies, encompassing fully 84% of the examinee population. An information function as flat as this is unusual. It represents a real success from a test design point of view, for it means that a very large proportion of examinees are tested with equal accuracy. Typically, information functions for fixed format tests peak in the middle and taper off quickly on both sides. Only adaptive tests (Wainer, Dorans et al. 2000), which are individually constructed to be optimal for each examinee, can be counted on to yield information curves like this one on a regular basis. Using test information (or its inverse, the standard error) as a representation of

accuracy is likely to be far more useful than a single reliability statistic.

INSERT FIGURE 5 HERE

Because such a high, flat test information function is so unusual, a second question immediately comes to mind. What constitutes such a curve? It might have come about in many ways. There may have been 12 highly peaked curves, which when summed, yielded the agreeable flat function seen in Figure 5. Or there might have been just a couple of very wonderful items that yielded this curve and the others may be essentially worthless. But the correct answer, easily seen by plotting individual information curves (Figure 6) obtained from Equation (8), is that all items share the same overall information structure, although the information for items 1 and 2 (the two bottom curves) is somewhat less than the others. A comparison with the analogous plot from the North Carolina Test of Computer Skills (Figure 2) shows marked differences. In Figure 2, all items were mostly informative, and differentially so across ability levels.

INSERT FIGURE 6 HERE

It is hoped that this brief example provides an illustration of what help this scoring model can provide even when the test does not require its full power. That is, even in cases where the variance of the testlet effects is negligible, our approach yields a number of benefits. First, it allows one to coherently combine information from items of varying design, thus providing an accurate assessment of test information. Second, MCMC draws facilitate straightforward computation of the posterior distribution of standard quantities of interest. And third, our approach allows one to treat items as independent with "confidence," as the assumption of CI has been empirically verified.

6 Conclusions

The North Carolina Test of Computer Skills and the TSE are testlet-based tests in which at least some of the items are polytomously scored. While there are psychometric models that can fit tests made up of polytomous items (e.g., Samejima, 1969; Bock, 1972), there are no psychometric models currently available that can accommodate such tests when within-testlet local dependence is likely. In both our simulations and our analysis of real data, we have shown how this model can be used to score such tests and provide estimates of test precision that are neither as optimistic as models that incorrectly assume conditional independence nor as pessimistic as those that only use total score. Furthermore, we have shown that in some cases testlet structures yield local dependence, while in other cases they yield none. As mentioned earlier, examining predictors of this sort will likely be of great practical interest.

There are currently two trends in modern testing. The first is a movement away from what is viewed as the atomistic nature of discrete multiple-choice items and toward the use of testlets as a way of providing context. The second is toward computerizing tests, both to allow the testing of constructs difficult or impossible to test otherwise, and to improve the efficiency of tests by making content adaptive to individal proficiency. Adaptive tests are often engineered to stop after measuring examinee proficiency to a predetermined level of precision. The model that we have proposed and tested here, by allowing the inclusion of testlets scored in a variety of ways, and therefore providing accurate assessments of information, should prove to be a useful complement to these modern trends.

Our model can have immediate application in many testing programs, but its principal value lies in the future. It provides a mechanism for scoring a test that is made up of any combination or grouping of multiple-choice items, fill-in-the-blank items, and other constructed-response items

that are judged by expert raters. This means that test developers can focus on what test structure measures best the construct of interest and trust that there will be an appropriate scoring model available for it. This has not been the case in the past.

Our analyses of the North Carolina Test of Computer Skills show one situation in which this testlet scoring model is required. The analysis of the TSE exam provides assurance that current test scoring methods are adequate. But there are also other instances in which this model can be profitably and immediately employed. For example, consider the Graduate Record Examinations (GRE) Writing Assessment, in which examinee writing samples are scored by two judges. If those judges disagree strongly, a third judge–a "master rater"–is brought in. The examinee's final score is the average of the master rater's judgment and the judge that is closest to the master rater. This approach is wasteful of expert judgment. We can be more efficient by using the new model. Let us think of the essay and its associated polytomous ratings as a single testlet in which the judges are the "items." We can then leave in all of the raters and allow the model to assign weights to the raters that are proportional to the extent to which each rater's judgment is related to the common underlying trait. This is theoretically akin to the observed score approach espoused by Braun and Wainer (1989).

Because this model is so general, its potential uses are very broad indeed. Although the requirements implicit in building testlets adaptively were the driving forces behind the development of the model, that is only one of its areas of application. Others will emerge as test developers' eyes grow accustomed to the light that is shed by this approach.

Perhaps the greatest value of this model is the role that it will play as the foundation of a general educational diagnostic system. As the model is currently configured, it can answer many "how much" questions: How much ability does this examinee have? How much difficulty does this item have? How much local dependence exists within this testlet? These are often important

questions, but for diagnosis as well as scientific understanding, we need answers to a similar set of "why" questions: Why does this examinee show this much ability? Why is this item so difficult? Why does this testlet exhibit so much local dependence? To answer such questions, one further generalization in which each of the parameters of interest is decomposed into a function of covariates is required. The developments discussed here are necessary precursors to this ultimate goal.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88,* 669-679.

Baker, F. B. (1992). *Item response theory.* New York, NY: Marcel-Dekker Inc.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29-51.

Bradlow, E. T. & Wainer, H. (1998). Some statistical and logical considerations when rescoring tests. *Statistica Sinica, 8,* 713-728.

Bradlow, E. T., Wainer, H., & Wang, X (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Bradlow, E. T., & Zaslavsky, A. M. (1999). A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *Journal of the American Statistical Association, 94,* (445), 43-52.

Braun, H. & Wainer, H. (1989). Making essay test scores fairer with statistics. In J. Tanur et al. (Eds.), *Statistics: A guide to the unknown* (3rd ed., pp. 178-187). San Francisco, CA: Holden Day.

Drasgow, F., Levine, M. A., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *The British Journal of Mathematical Psychology, 38,* 67-86.

Educational Testing Service (1995). *TSE score user's manual.* Princeton, NJ: Educational Testing Service.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association, 85,* 398-409.

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

Gelman, A., Roberts, G, & Gilks, W. (1995). Efficient metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 5.* New York, NY: Oxford University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science, 7,* 457-511.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 54,* 93-108.

Levine, M. V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53,* 161-176.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, 7.*

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

Roberts, G. O., & Smith, A. F. M. (1993). Bayesian computation via the Gibbs Sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, 55,* 3-23.

Rosa, K., Swygert, K., Nelson, L. & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed response items: Scale scores for patterns of summed scores. D. Thissen and H. Wainer (Eds.), *Test scoring* (pp. 253-292). Hillsdale, NJ: Lawrence Erlbaum.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs, 17.*

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589-617.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82,* 528-540.

Thissen, D. (1991). *MULTILOG user's guide* (version 6). Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L. & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement 26,* 247-260.

Thissen, D. & Wainer, H. (Eds.) (2001). *Test Scoring.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Wainer, H., Bradlow, E. T., & Du, Z (2000). Testlet response theory. An analog for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.),

Computerized adaptive testing: Theory and practice. Kluwer-Nijhoff, 245-270.

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-202.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* (2), 157-187.

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

Zhang, J. & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213-249.

Appendix

To implement the Gibbs sampler for our model, we need to sample from the set of conditional distributions corresponding to parameters $a_j, b_j, c_j, g_r, \theta_i, \gamma_{id(j)}, \mu_a, \mu_b, \mu_c, \sigma_a^2, \sigma_b^2, \sigma_c^2, \sigma_{d(j)}^2$. However, these conditional distributions can be categorized into a smaller number of types due to the similar structure that exists across parameters. For each type, we write down one of the forms (the others follow directly). Specifically, all of the parameters' conditional distributions are of the type:

(1) Abilities θ_i,

$$[\,\theta_i \mid Y, \Lambda_{-\theta_i}\,] \sim \prod_{j\in\mathcal{J}_1}\left[\frac{\frac{e^{q_j}}{1+e^{q_j}}+e^{a_j(\theta_i-b_j-\gamma_{ik(j)})}}{1+e^{a_j(\theta_i-b_j-\gamma_{ik(j)})}}\right]^{y_{ij}}\left[\frac{1-\frac{e^{q_j}}{1+e^{q_j}}}{1+e^{a_j(\theta_i-b_j-\gamma_{ik(j)})}}\right]^{1-y_{ij}} \cdot \prod_{j\in\mathcal{J}_2}\prod_{r=1}^{R_j}(\Phi(g_r-t_{ij})-\Phi(g_{r-1}-t_{ij}))^{1(v_{ij}=r)} \cdot e^{-\frac{1}{2}\theta_i^2}, \text{ for } i=1,\ldots,I,$$

where $1_{(\cdot)}$ denotes an indicator function and t_{ij} is the linear predictor of score as in (4).

(2) Item discriminations a_j, item difficulties b_j, and transformed guessing parameters for the binary items $q_j = \text{logit}(c_j)$,

$$[\,a_j \mid Y, \Lambda_{-a_j}\,] \sim \prod_{i=1}^{I}\left[\frac{\frac{e^{q_j}}{1+e^{q_j}}+e^{a_j(\theta_i-b_j-\gamma_{ik(j)})}}{1+e^{a_j(\theta_i-b_j-\gamma_{ik(j)})}}\right]^{y_{ij}}\left[\frac{1-\frac{e^{q_j}}{1+e^{q_j}}}{1+e^{a_j(\theta_i-b_j-\gamma_{ik(j)})}}\right]^{1-y_{ij}} \cdot e^{-\frac{1}{2\sigma_a^2}(a_j-\mu_a)^2} \text{ for every } j\in\mathcal{J}_1,$$

(3) Item discriminations a_j and item difficulties b_j for the polytomous items,

$$[\,a_j \mid Y, \Lambda_{-a_j}\,] \sim \prod_{i=1}^{I}\left[\prod_{r=1}^{R_j}(\Phi(g_r-\mu_{ij})-\Phi(g_{r-1}-\mu_{ij}))^{1(v_{ij}=r)}\right] \cdot e^{-\frac{1}{2\sigma_a^2}(a_j-\mu_a)^2}, \text{ for every } j\in\mathcal{J}_2,$$

(4) Testlet effects $\gamma_{id(j)}$,

$$
\begin{aligned}
&[\gamma_{ik} \mid Y, \Lambda_{-\gamma_{ik}}] \sim \\
&\prod_{j \in \mathcal{K}_i \cap \mathcal{J}_1} \left[\frac{\frac{e^{q_j}}{1+e^{q_j}} + e^{a_j(\theta_i - b_j - \gamma_{ik(j)})}}{1 + e^{a_j(\theta_i - b_j - \gamma_{ik(j)})}} \right]^{y_{ij}} \left[\frac{1 - \frac{e^{q_j}}{1+e^{q_j}}}{1 + e^{a_j(\theta_i - b_j - \gamma_{ik(j)})}} \right]^{1-y_{ij}} \\
&\cdot \prod_{j \in \mathcal{K}_k \cap \mathcal{J}_2} \left[\prod_{r=1}^{R_j} \left(\Phi(g_r - t_{ij}) - \Phi(g_{r-1} - t_{ij}) \right)^{1(y_{ij}=r)} \right] \\
&\cdot e^{-\frac{1}{2\sigma^2_{\gamma(k)}} \gamma^2_{ik}} \quad \text{for every } i = 1, \ldots, I, \text{ and } k = 1, \ldots, K,
\end{aligned}
$$

(5) Polytomous item cutoffs,

$$\text{Uniform}(\max\{t_{ij} : y_{ij} = r\}, \min\{t_{ij} : Y_{ij} = r+1\}),$$

uniform random variable.

(6) Prior means μ_a, μ_b, μ_q,

$$[\mu_a \mid Y, \Lambda_{-\mu_a}] \propto e^{-\frac{1}{2\sigma^2_a}(\mu_a - \bar{a})^2}, \text{ where } \bar{a} = \frac{1}{J} \sum_{j=1}^{J} a_j,$$

a normal random variable.

(7) Prior variances $\sigma^2_a, \sigma^2_b, \sigma^2_q, \sigma^2_{d(k)}$,

$$\left[\sigma^2_a \mid Y, \Lambda_{-\sigma^2_a} \right] \sim \sigma_a^{-2\left[\left(\frac{J}{2} + \frac{1}{4}\right) - 1\right]} e^{-\sigma_a^{-2}\left[\frac{1}{2} + \frac{1}{2}\sum_{j=1}^{J}(a_j - \mu_a)^2\right]},$$

an inverse-gamma random variable.

As we note in text for the current study, distributions for forms (1), (2), (3), and (4) were obtained using a Metropolis-Hastings step. Draws from (5), (6), and (7) were obtained directly from uniform, normal, and inverse-gamma distributions respectively. Computation for each iteration of the sampler took 1/3 second for simulated data consisting of 1,000 simulees, and roughly 1.5 seconds per iteration for real data examples run on a Pentium 3, 550 MHZ computer programmed in C.

Table 2: Correlations Between True Parameters and Estimated Posterior Means.

Simulation Number	Number of Categories	Number of Testlets	var(γ)	a	b	c	g_r	θ
1	2	6	0.0	0.891 (0.051)	0.990 (0.003)	0.635 (0.101)	NA	0.924 (0.005)
2	5	6	0.5	0.886 (0.036)	0.993 (0.003)	0.660 (0.123)	0.972 (0.013)	0.947 (0.006)
3	10	6	1.0	0.846 (0.093)	0.993 (0.004)	0.565 (0.134)	0.973 (0.006)	0.934 (0.005)
4	2	3	0.5	0.832 (0.054)	0.992 (0.002)	0.645 (0.098)	NA	0.911 (0.003)
5	5	3	1.0	0.891 (0.063)	0.995 (0.002)	0.600 (0.140)	0.984 (0.007)	0.902 (0.015)
6	10	3	0.0	0.951 (0.013)	0.994 (0.002)	0.619 (0.128)	0.986 (0.007)	0.979 (0.004)
7	2	2	1.0	0.870 (0.054)	0.984 (0.003)	0.500 (0.103)	NA	0.884 (0.015)
8	5	2	0.0	0.927 (0.051)	0.992 (0.004)	0.633 (0.208)	0.979 (0.007)	0.972 (0.005)
9	10	2	0.5	0.939 (0.010)	0.995 (0.002)	0.536 (0.073)	0.986 (0.007)	0.916 (0.009)

Note: Reported values are the average over five replicated data sets. Values in parentheses are corresponding standard deviations. NA values indicate those cases in which all items were binary and hence cutoffs did not have to be estimated.

Table 3: Mean Square Error Between True Parameters and Estimated Posterior Means.

Simulation Number	Number of Categories	Number of Testlets	var(γ)	a	b	c	g_r	θ
1	2	6	0.0	1.54 (0.31)	0.75 (0.33)	5.52 (1.38)	NA	0.14 (0.01)
2	5	6	0.5	1.46 (0.48)	0.46 (0.13)	3.50 (1.04)	0.96 (0.52)	0.11 (0.01)
3	10	6	1.0	2.29 (1.19)	0.52 (0.18)	6.12 (3.04)	2.34 (0.70)	0.13 (0.01)
4	2	3	0.5	1.47 (0.20)	0.68 (0.31)	5.25 (1.46)	NA	0.18 (0.01)
5	5	3	1.0	1.20 (0.53)	0.34 (0.12)	4.00 (1.56)	0.48 (0.19)	0.18 (0.02)
6	10	3	0.0	0.79 (0.14)	0.50 (0.36)	3.84 (1.74)	1.17 (0.50)	0.04 (0.01)
7	2	2	1.0	1.80 (0.41)	1.29 (0.55)	9.87 (5.22)	NA	0.21 (0.02)
8	5	2	0.0	0.97 (0.46)	0.55 (0.30)	4.57 (2.59)	0.63 (0.15)	0.06 (0.01)
9	10	2	0.5	0.92 (0.36)	0.48 (0.17)	4.94 (2.16)	1.19 (0.46)	0.16 (0.02)

Note: Reported values are the average over five replicated data sets. Values in parenthesis are corresponding standard deviations. All table values are multiplied by 10^2. NA values indicate those cases in which all items were binary and hence cutoffs did not have to be estimated.

Table 4: Correlations between True Value and Posterior Mean Slope Parameter (a).

	Number of testlets			
Number of categories	6	3	2	Mean
2	0.89	0.83	0.87	0.86
5	0.89	0.89	0.93	0.90
10	0.85	0.95	0.94	0.91
Mean	0.87	0.89	0.91	0.89

Table 5: Correlations Between True Value and Posterior Mean Difficulty Parameter (b).

	Number of testlets			
Number of categories	6	3	2	Mean
2	0.990	0.992	0.984	0.989
5	0.993	0.995	0.992	0.993
10	0.993	0.994	0.995	0.994
Mean	0.992	0.994	0.990	0.992

Table 6: Correlation Between True Value and Posterior Mean for Each Parameter.

	Var(γ)		
Parameters	0.0	0.5	1.0
a	0.92	0.89	0.87
b	0.99	0.99	0.99
c	0.60	0.42	0.55
d	0.98	0.98	0.98
θ	0.96	0.92	0.91

Table 7: Results From Testlet Model Fitted to the North Carolina Test of Computer Skills.

	Number of polytomous items	Number of dichotomous items	Total items	Testlet effects var(γ)
Keyboarding	3	0	3	0.03
Word processing/editing	0	10	10	2.80
Database use	3	4	7	0.78
Spreadsheet use	1	5	6	2.58
Total	7	19	26	

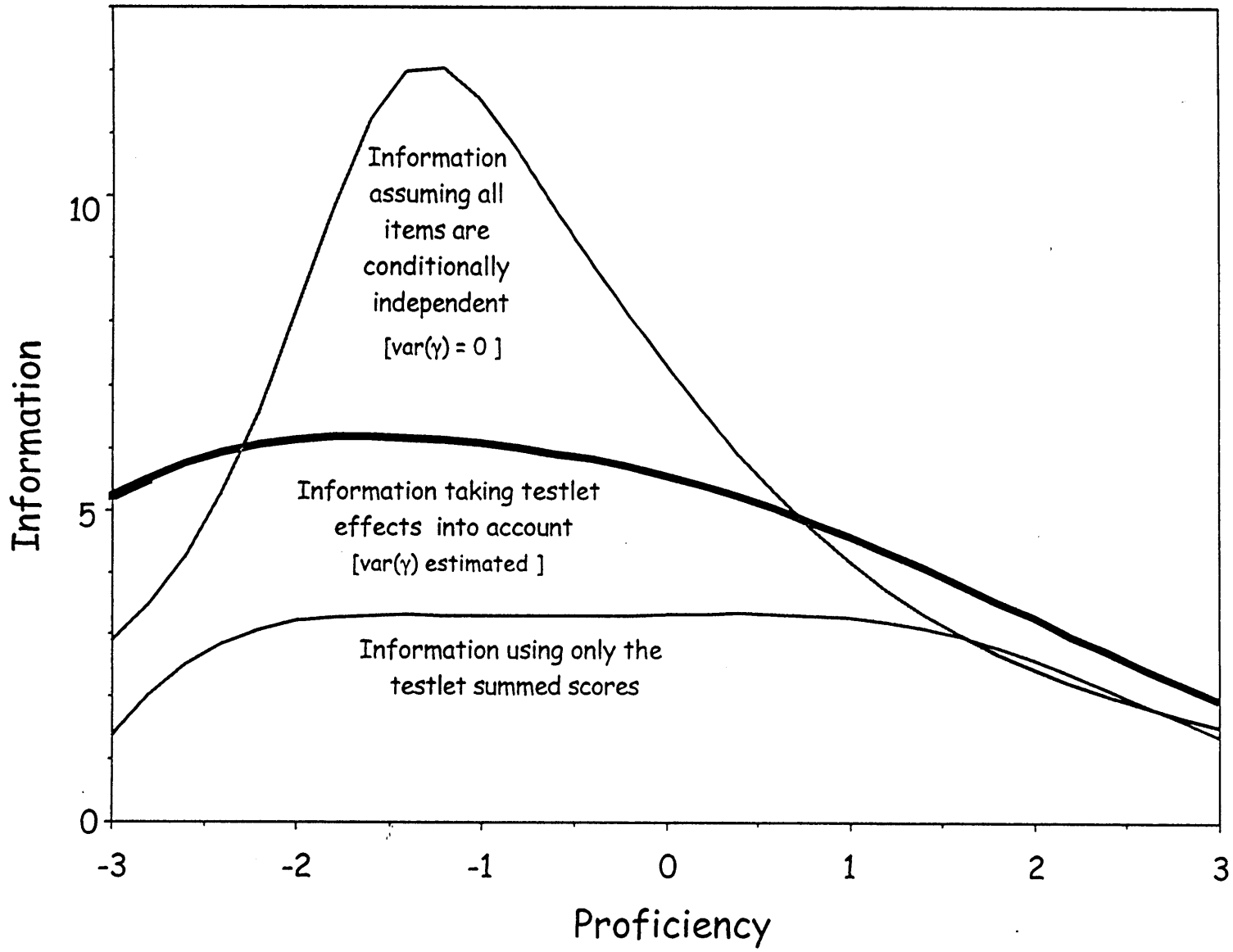

Figure 1. Information functions for the performance sections of the North Carolina Test of Computer Skills estimated in three ways.

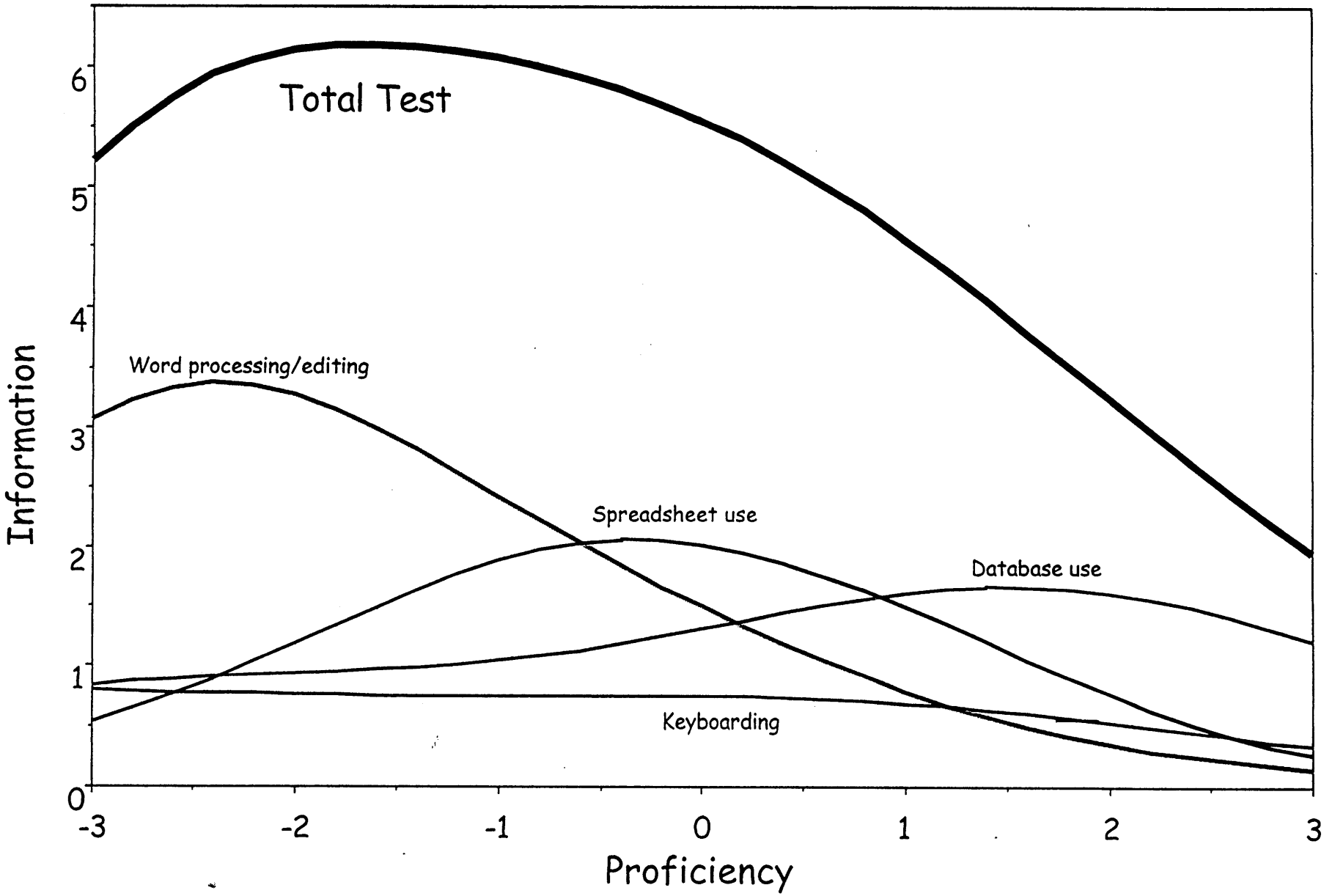

Figure 2. Information from the performance sections of the North Carolina Computer Skills Test broken down by testlet and showing how each subject area spans a specific proficiency region.

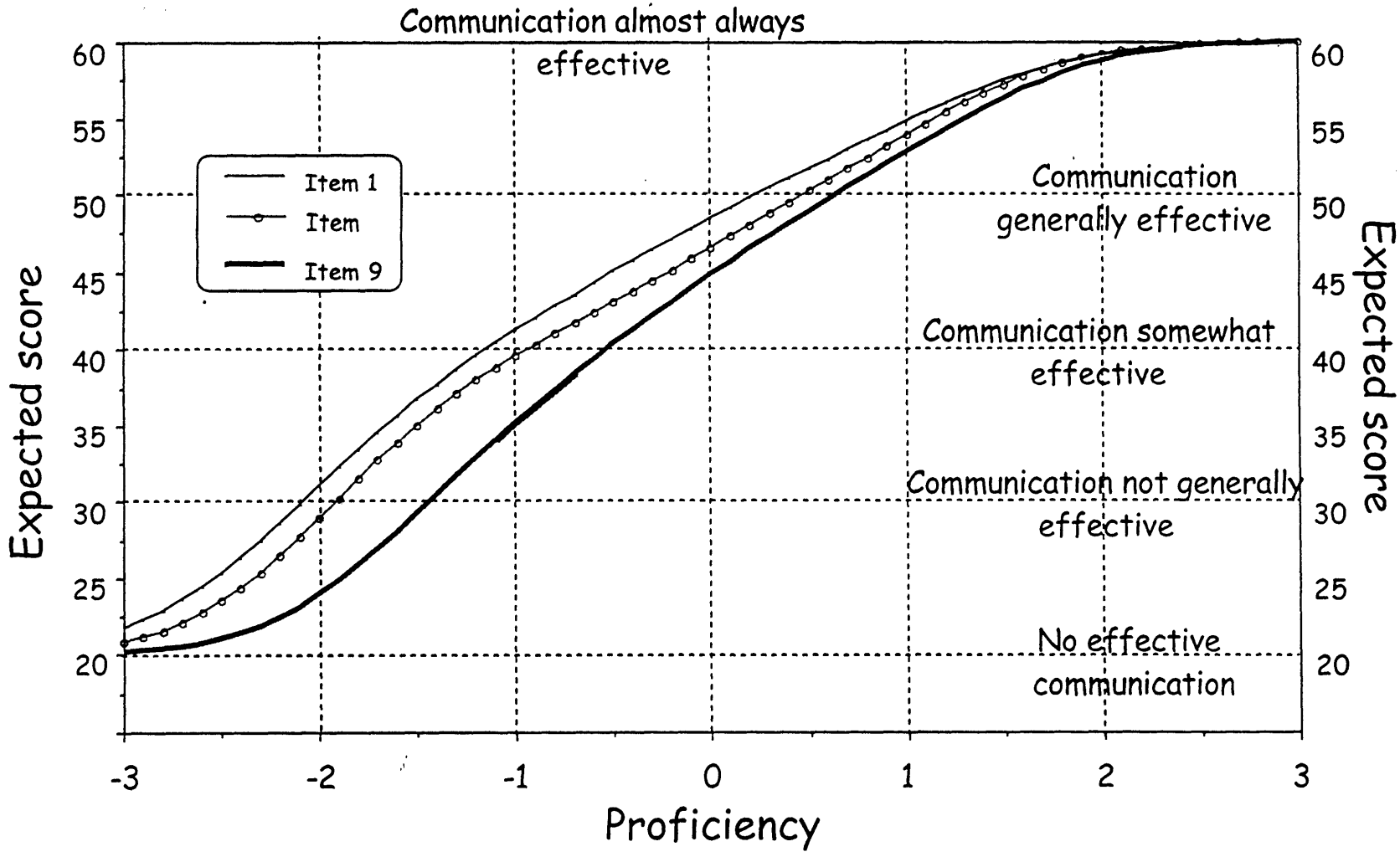

Figure 3. Expected score curves for three TSE items

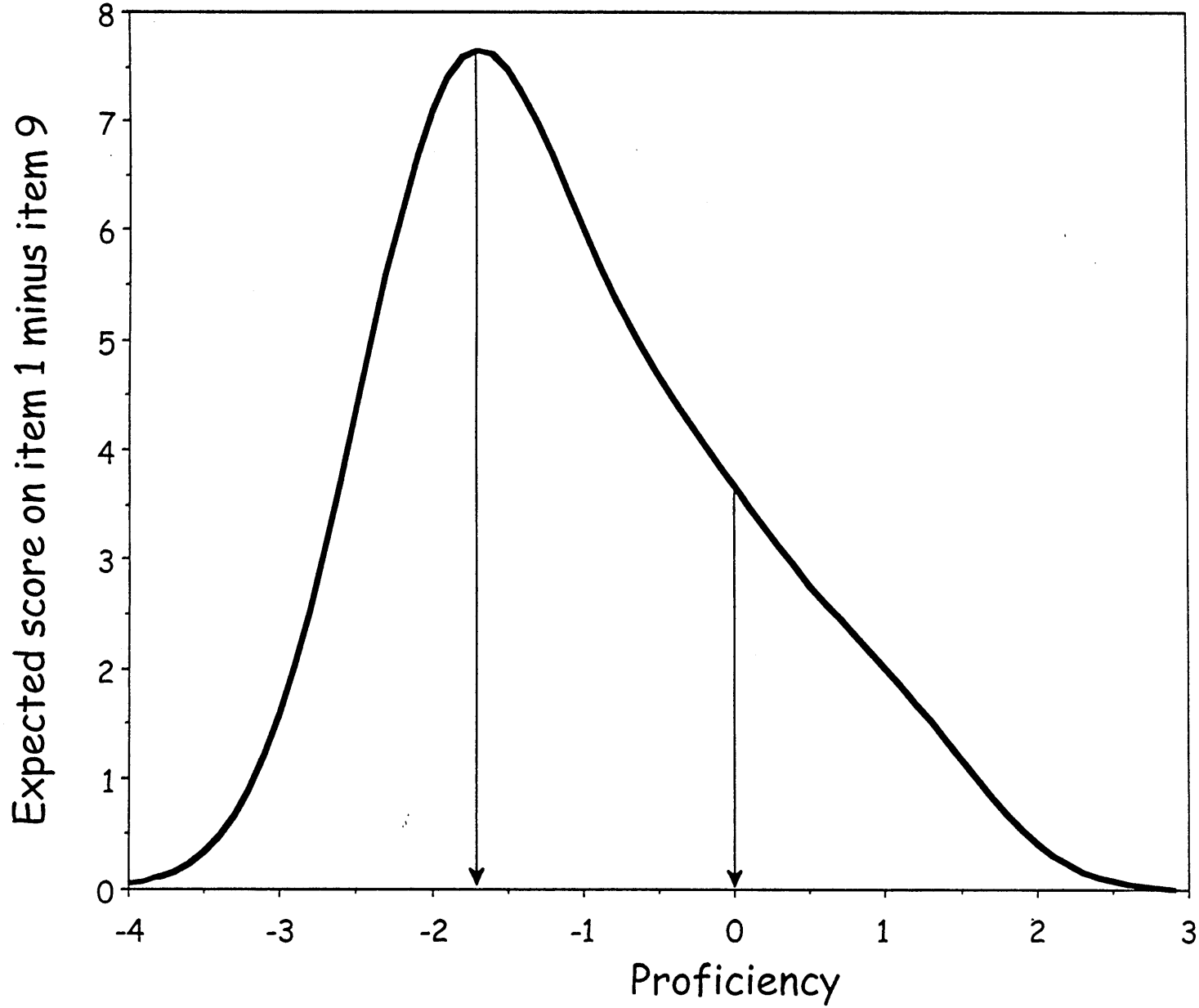

Figure 4. The difference in expected score between the hardest and the easiest TSE items.

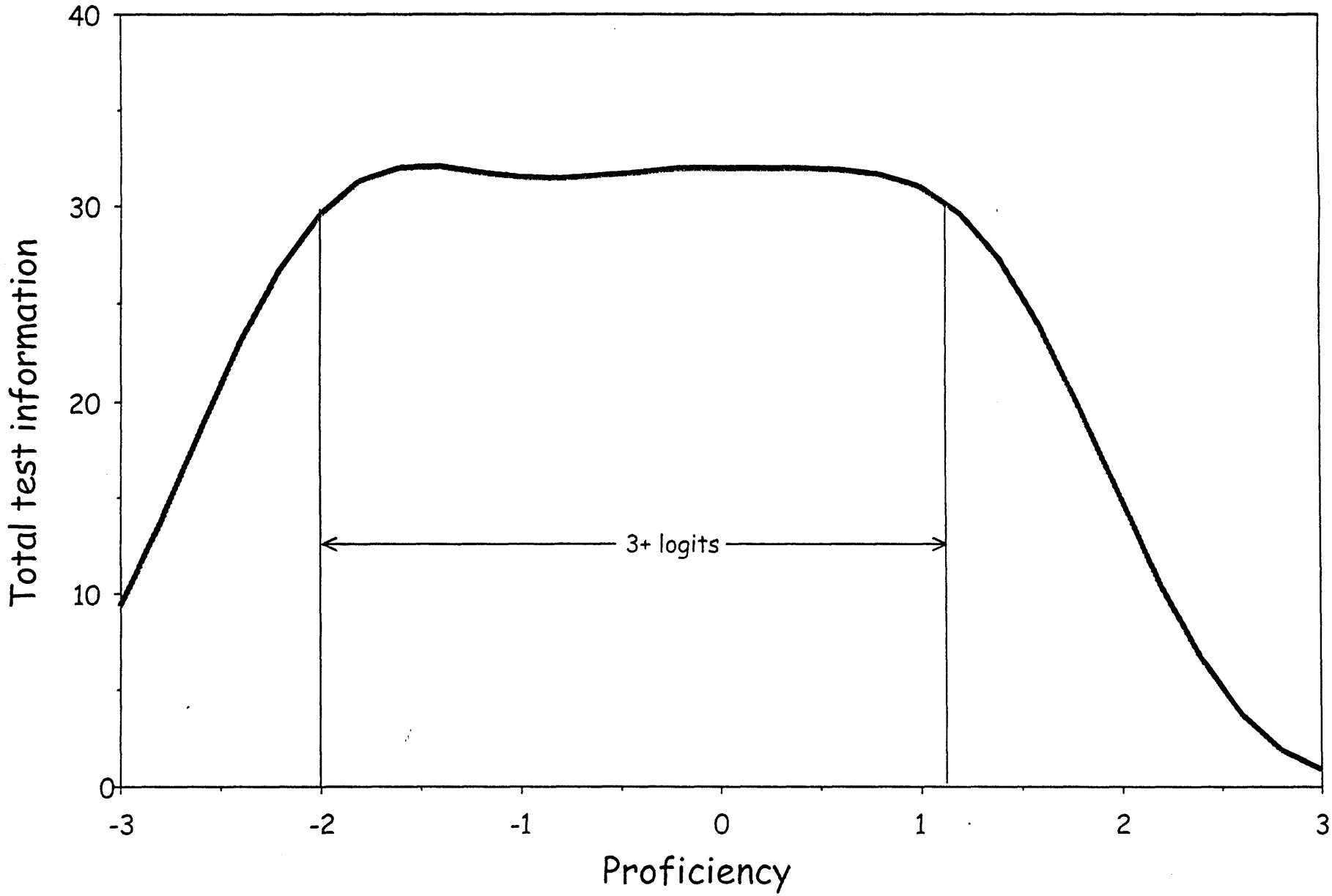

Figure 5. The total test information function for TSE shows that the test provides equally good estimation over a very wide range of proficiencies.

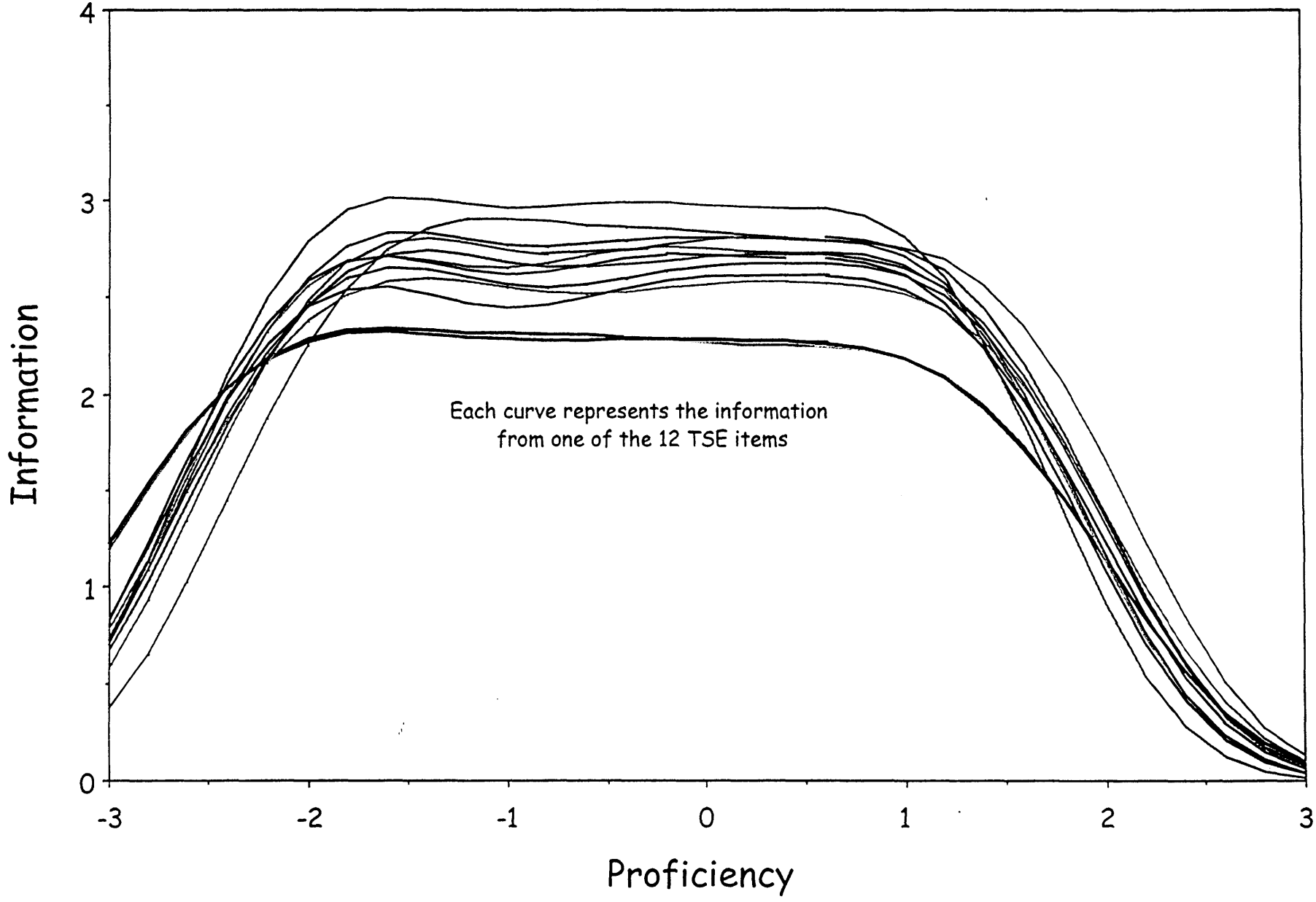

Figure 6. Information functions for all items of the TSE; all items provide uniformly high information over the proficiency band of interest.