DOCUMENT RESUME

ED 333 034 TM 016 542

AUTHOR	Junker, Brian W.; And Others
TITLE	Structural Robustness and Ability Estimation in Item Response Theory: A Survey.
SPONS AGENCY	National Inst. of Mencal Health (DHHS), Bethesda, Md.; Office of Naval Research, Arlington, Va.
PUB DATE	Apr 91
CONTRACT	NIMH-MH15758; ONR-N00014-90-J-1984; ONR-N00014-91-J-1208
NOTE	35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE	Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *Equations (Mathematics); *Estimation (Mathematics); *Item Response Theory; Literature Reviews; *Mathematical Models; Maximum Likelihood Statistics; *Robustness (Statistics); Theory Practice Relationship
IDENTIFIERS	Ability Estimates; *Local Independence (Tests); Unidimensionality (Tests)

ABSTRACT

Some item response theory (IRT) techniques work in applications even though the usual structural IRT assumptions, and local independence (LI) in particular, do not hold. When the departure from LI is too great, traditional procedures will break down. Although violations of strictly unidimensional, monotone, locally independent latent structure can sometimes be modeled and exploited, many situations call for a unidimensional approach that is tolerant of minor violations of strict unidimensionality. Departures from strict unidimensionality can be detected, and the influence of these departures on a variety of LI-based ability estimators can be measured. A convenient universe of models near the LI model in which to investigate structural robustness issues is provided by the essential unidimensionality modeling approach of W. F. Stout (1990). Theoretical results underpinning the approach are surveyed. Work in progress to apply these results in practical settings is described; the goal is to develop guidelines for the detection of departures from unidimensionality. Three data tables and 10 figures illustrate the discussion. A 33-item list of references is included. (Author/SLD)

Structural Robustness and Ability Estimation in Item Response Theory: A Survey

Brian W. Junker*
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
(412) 268-8873
brian@stat.cmu.edu

A paper presented at the Annual Meeting of the American Educational Research Association, Chicago Illinois, April 3, 1991.

Abstract

Some item response theory (IRT) techniques "work" in applications even though the usual structural IRT assumptions, and local independence (LI) in particular, do not hold. When the departure from local independence is too great, traditional procedures will break down. Although violations of strictly unidimensional, montone, locally independent latent structure can sometimes be modeled and exploited, many situations call for a unidimensional approach that is tolerant of minor violations of strict unidimensionality (e.g., Drasgow and Parsons, 1983; Spray and Ackerman, 1987; Reckase 1990). Departures from strict unidimensionality can be detected, and the influence of these departures on a variety of LI-based ability estimators can be measured. A convenient universe of models *near* the LI model in which to investigate structural robustness issues is provided by Stout (1990b)'s *essential unidimensionality* modeling approach. In this paper we survey theoretical results underpinning this approach, and report on work in progress to apply these results in practical settings.

Keywords: Essential independence, wrong-model analysis, local dependence, maximum likelihood, information, posterior ability distributions, model-fit indices.

*Research supported in part by grants ONR-N00014-90-J-1984, ONR-N00014-91-J-1208, and NIMH-MH15758. Some of the work reported here is joint with coauthors Bertrand S. Clarke (Purdue University) and William F. Stout (University of Illinois).

Contents

1 Int̃roduction

Traditional unidimensional monotonic Item Response Theory (IRT) provides a useful but overly simple model of examinees' responses to standardized test questions. For example, Drasgow and Parsons (1983) assess the shortcomings of the LI-based unidimensionality approach to IRT as follows:

> One way in which most current item response theories (IRTs) are surely incorrect is in their assumption of a unidimensional latent trait space ... [I]t seems clear that researchers should be more concerned with the robustness of estimation techniques to minor violations of dimensionality assumptions than with the possibly never-ending task of measuring all latent variables that underlie responses in a particular content domain.

We are compelled to understand this *structural robustness* question because it is central to current IRT practice. It is widely accepted that the traditional IRT models do not exactly reflect the item response process; yet because the traditional inference procedures (in the form of computer programs such as LOGIST and BILOG) are so accessible, traditional IRT is applied to item response data anyway. Can we trust the inferences from these misspecified models?

Although violations of **strictly unidimensional**, $d_L = 1$, structure—i.e., models satisfying the stronger traditional assumptions of *local independence (LI)* and *monotonicity (M)*—can sometimes be modeled and exploited (whether by introducing new "dependence" parameters as in Jannarone (1986) and Gibbons, Bock and Hedeker (1989), or by an explicitly multidimensional approach as that of Reckase (1990)), many situations call for a unidimensional approach that is tolerant of minor violations of strict unidimensionality (e.g., Drasgow and Parsons, 1983; Spray and Ackerman, 1987; Reckase 1990). It is also important to note that an "acceptable" level of departure from strict unidimensionality may depend on the particular application; for example, ability rank estimates on a particular section of the Graduate Record Examination may be more tolerant to violations of unidimensionality than detailed item analysis of the same items.

Departures from strict unidimensionality can be detected, and the influence of these departures on a variety of LI-based ability estimators can be measured. In this paper we survey theoretical results underpinning this approach, and report on work in progress to apply these results in practical settings. A convenient universe of models *near* the LI model in which to investigate structural robustness issues is provided by Stout (1990b)'s **essential unidimensionality**, $d_E = 1$, modeling approach. The main ideas of essential independence, summarized in Section 2, are due to Stout (1987, 1990). The approach to structural robustness for maximum likelihood estimation of ability outlined in Section 3 is due to Junker (1991b). The more general statistical considerations of Sections 4 and 5 represents the joint work of Clarke and Junker (1991). Finally the work on two new indices of unidimensionality described in Sections 6 and 7 represent ongoing joint work of Junker and Stout. Owing to the "survey" nature of this paper, once Section 2 is read the remaining sections may be read in any order.

2 Essential independence

A successful approach to identifying unidimensional latent structure outside the strict LI/M framework has been pursued in the seminal work of Stout (1987) and Stout (1990b), and extended by Junker (1988) and Junker (1991b). The main idea, which borrows from both the "large sample theory" tradition in mathematical statistics and the "factor analysis" tradition in psychometrics, is that of *essential independence*.

For any (infinite) sequence of items $\underline{X} = (X_1, X_2, X_3, \ldots)$ (dichotomous or polytomous), we define *bounded item scores* to be functions $A_j(X_j)$ such that $\exists M < \infty$ such that $|A_j(X_j)| \leq M \; \forall j$. In the special case that each X_j takes on ordered, discrete values $\xi_{j1} \leq \xi_{j2} \leq \ldots$, we will call a bounded item score an *ordered item score* if moreover $A_j(\xi_{jk}) \leq A_j(\xi_{j(k+1)}) \; \forall k$ (in the dichotomous case, $A_j(0) \leq A_j(1)$, for example). Also, we define a *bounded test score* to be the average of the first J bounded item scores $\overline{A}_J = \frac{1}{J}\sum_{j=1}^{J} A_j(X_j)$.

Definition 2.1 *The infinite item sequence $\underline{X}$ is* essentially independent (EI) with respect to $\underline{\Theta}$ *if and only if*

$$\lim_{J\to\infty} \mathrm{Var}(\overline{A}_J | \underline{\Theta} = \underline{\theta}) = 0 \tag{1}$$

for all bounded test scores $\overline{A}_J$.

It can be shown that therefore $\overline{A}_J$ is a consistent estimator of the "true score" $\overline{A}_J(\underline{\theta}) = E[\overline{A}_J | \underline{\Theta} = \underline{\theta}]$, as $J \to \infty$. In particular, for a sequence of dichotomous items $\underline{X}$, EI implies that the *proportion correct* score $\overline{X}_J$ consistently estimate values of the the *test characteristic curve* (TCC) $\overline{P}_J(\underline{\theta})$, as $J \to \infty$.

Definition 2.2 *The item sequence $\underline{X}$ is* essentially unidimensional, $d_E = 1$, *if and only if*

> **EI:** *$\underline{X}$ is EI with respect to a unidimensional Θ; and*
>
> **LAD**[1]: *for every set of ordered item scores, the "true score" $\overline{A}_J(\theta)$ is nondecreasing in θ.*

If no such unidimensional Θ exists, we write $d_E > 1$.

All of the theoretical results in this paper apply to polytomously-scored—and in some cases continuously-scored—items, but for simplicity we will focus mostly on the familiar dichotomous case in which $X_j = 0$ or 1.

When an item sequence has $d_E = 1$ the true score $\overline{A}_J(\theta)$ may be estimated with $\overline{A}_J$ and then inverted to produce estimates $\tilde{\theta} = \overline{A}_J^{-1}(\overline{A}_J)$ of θ itself; in particular, $\overline{P}_J^{-1}(\overline{X}_J) \xrightarrow{p} \theta$ as $J \to \infty$. The notion of *essential unidimensionality*, $d_E = 1$, should be contrasted with *strict unidimensionality*, $d_L = 1$, under which both local independence (LI) and monotone (M) increasing ICC's are required. In particular, any model satisfying LI also satisfies EI. See Stout (1990b) and Junker (1991b) for details.

[1] "LAD" stands for *locally asymptotically discriminating*. There is a technical detail about the possibility of "flat" true scores which need not concern us here. See Junker (1991b) or Stout (1990b) for details.

The $d_E = 1$ condition is built out of the good θ-estimation properties of total test scores under strict unidimensionality. The EI condition is explicitly designed to characterize unidimensional behavior—the "driving" of $\overline{X}_J$ by a single dominant trait θ—under conditions other than local independence. Nandakumar and Stout (Stout, 1987; Nandakumar, 1987, 1989, 1991a, 1991b) have investigated the practical assessment of essential unidimensionality in a variety of finite-length tests with minor violations of the $d_L = 1$ representation.

3 Structural robustness and maximum likelihood

Stout (1990b)'s definition of essential independence covers a broad range of situations in which one might wish to assert that the fundamental behavior of the data is unidimensional, although "nuisance traits" prevent strict $d_L = 1$ from holding. The influences of these nuisance traits are sufficiently small that it is tempting to use strictly unidimensional estimation techniques to estimate the dominant (and interesting) trait, rather than to take an explicitly multitrait approach. For example, maximum likelihood estimation of the dominant or target θ can be examined in this light.

In the binary (dichotomous) case, in which X_j takes the value 0 or 1 depending on the examinee's answer to the j^{th} item, the $d_L = 1$ likelihood is

$$P[\underline{X}_J = \underline{x}_J \mid \Theta = \theta] = \prod_{j=1}^{J} P_j(\theta)^{x_j}[1 - P_j(\theta)]^{1-x_j}, \tag{2}$$

with monotone item characteristic curves (ICC's) $P_j(\theta) = P[X_j = 1 \mid \Theta = \theta]$. If the log-likelihood is sufficiently smooth, the MLE must solve the likelihood equation

$$0 \equiv L'_J(\hat{\theta}_J) = \sum_{j=1}^{J} \lambda'_j(\hat{\theta}_J)[X_j - P_j(\hat{\theta}_J)], \tag{3}$$

where $\lambda_j(\theta) = \log P_j(\theta)/(1 - P_j(\theta))$ It should be noted that the "minimum discrimination" condition LAD in Definition 2.2 plays a crucial role in the rigorous proof of consistency of $\hat{\theta}_J$: LAD guarantees that the average information function

$$\bar{I}_J(\theta) = \frac{1}{J}\sum_{j=1}^{J} \lambda'_j(\theta)P'_j(\theta) \geq \epsilon_\theta > 0, \tag{4}$$

as $J \to \infty$. See Junker (1991b) for details.

Theorem 3.1 *Let $\underline{X}$ be a dichotomous item sequence with sufficiently smooth ICC's satisfying EI and (4). Then there exists a sequence $\{\hat{\theta}_J : J \geq J_0\}$ of roots of (3) such that*

$$\lim_{J\to\infty} P[\,|\,\hat{\theta}_J - \theta\,| < \epsilon \mid \Theta = \theta] = 1, \tag{5}$$

for every $\epsilon > 0$ (i.e., $\hat{\theta}_J \xrightarrow{P} \theta$, given $\Theta = \theta$, as $J \to \infty$).

Indeed, (3) may be expanded as

$$\begin{aligned}\frac{1}{J}L'_J(t) &= \frac{1}{J}L'(\theta) + \frac{1}{J}(t-\theta)L''_J(\theta) + \frac{1}{2J}L'''_J(\xi)\\ &= \frac{1}{J}\sum_{j=1}^{J}\lambda'_j(\theta)[X_j - P_j(\theta)] + (t-\theta)\left\{\frac{1}{J}\sum_{j=1}^{J}\lambda''_j(\theta)[X_j - P_j(\theta)] - \frac{1}{J}\sum_{j=1}^{J}\lambda'_j(\theta)P'_j(\theta)\right\}\\ &\quad +\frac{1}{2}(t-\theta)^2\frac{1}{J}L'''_J(\xi)\\ &= o_p(1) - (t-\theta)\left\{\bar{I}_J(\theta) + o_p(1) + O(t-\theta)\right\}.\end{aligned}$$

Note for example that, by Definition 2.1,

$$\frac{1}{J}\sum_{j=1}^{J}\lambda'_j(\theta)[X_j - P_j(\theta)] \xrightarrow{P} 0$$

as $J \to \infty$. The other terms are handled similarly.

In general there are problems with multiple roots of the likelihood equation (3) when the problem is set up in this manner; moreover it may be argued that a consistency result for a *theoretical* solution to (3) is of no value in practice. Fortunately, the same method of proof shows that the familiar practice of approximating a root of (3) by Newton's method still leads to consistent estimators, under EI.

Theorem 3.2 *Suppose the assumptions of Theorem 3.1 hold, and let $\tilde{\theta}_J$ be any sequence of consistent estimates of θ, given $\Theta = \theta$. Then the Newton's method improvement,*

$$\theta^*_J = \tilde{\theta}_J - \frac{\ell'_J(\tilde{\theta}_J)}{\ell''_J(\tilde{\theta}_J)},$$

is also consistent for θ.

Following the remarks after Definition 2.2, an obvious candidate for the initial guess in Theorem 3.2 is $\tilde{\theta}_J = \bar{P}_J^{-1}(\bar{X}_J)$.

In the usual LI ability estimation theory, we expect that the sequence $\hat{\theta}_J$ will be *asymptotically normal* and *efficient,*

$$J^{\frac{1}{2}}(\hat{\theta}_J - \theta) \sim AN(0, 1/\bar{I}_J(\theta)), \tag{6}$$

as $J \to \infty$, where $\bar{I}_J(\theta)$ is the traditional test information function introduced in (4). A result like (6) identifying the standard error of $\hat{\theta}_J$ is needed to do statistical inference using $\hat{\theta}_J$—or indeed, merely to know how well to trust $\hat{\theta}_J$ as an estimator of θ for particular fixed J that arise in applications. However, (6) may fail in the essentially unidimensional case in two interesting ways: it may be that asymptotic normality holds but the asymptotic variance is no longer $\bar{I}_J(\theta)^{-1}$; or it may be that asymptotic normality fails completely.

From (3) and the above results, we see that the consistency and asymptotic distribution of $\hat{\theta}_J$ is tied up with the behavior of the centered weighted averages

$$\begin{aligned} \frac{1}{J}\ell'_J(\theta) &= \overline{A}_J - \overline{A}_J(\theta) & (7) \\ &= \frac{1}{J}\sum_{j=1}^{J} a_j[X_j - P_j(\theta)], & (8) \end{aligned}$$

with $a_j \equiv \lambda'_j(\theta)$, where the dependence of a_j on θ does not matter since θ is fixed. Once again, let

$$\begin{aligned} \sigma_J^2(\theta) &= \mathrm{Var}(\overline{A}_J \mid \theta) \\ &= \frac{1}{J^2}\sum_{j=1}^{J} a_j^2 P_j(\theta)[1 - P_j(\theta)] + \frac{2}{J^2}\sum_{j=1}^{J}\sum_{J=1}^{i-1} a_i a_j \mathrm{Cov}(X_i, X_j \mid \Theta = \theta), \end{aligned}$$

and let

$$C_J(\theta) = \frac{2}{J}\sum\sum_{1 \le i < j \le J} \lambda'_i(\theta)\lambda'_j(\theta)\mathrm{Cov}\ (X_i, X_j | \theta). \qquad (9)$$

Theorem 3.3 *Suppose that the assumptions of Theorem 3.1 hold for the item sequence $\underline{X}$ and the latent trait Θ. Also suppose, given $\Theta = \theta$, that in (8),*

$$\frac{1}{\sigma_J(\theta)}[\overline{A}_J - \overline{A}_J(\theta)] \sim AN(0, 1). \qquad (10)$$

Finally, suppose $R(J)$ is a function for which $R^2(J)C_J(\theta)/J$ remains bounded. Then,

$$R(J)(\hat{\theta}_J - \theta) \sim AN\left(0, \frac{R^2(J)}{J}\frac{\overline{I}_J(\theta) + C_J(\theta)}{\overline{I}_J(\theta)^2}\right).$$

Moreover, if $\tilde{\theta}_J$ is any estimator for which $R(J)(\tilde{\theta}_J - \theta)$ is bounded in probability, θ_J^* from Theorem 3.2 is also asymptotically normal with the same asymptotic variance.

Theorems 3.1, 3.2 and 3.3 are structural robustness results: a method of estimating ability developed under $d_L = 1$ is robust to violations of $d_L = 1$ within the $d_E = 1$ framework, in the sense that it still converges to θ as test length J grows. However this robustness of consistency for $\hat{\theta}_J$ does not extend to robustness of variability. Nonefficient and non-normal asymptotic error distributions for $\hat{\theta}_J$ can be expected in many $d_E = 1$ situations; the deviation from the "efficient" LI-based standard error can be expressed in terms of the "index" $C_J(\theta)$. Further details, and extensions to the polytomous case, may be found in Junker (1991b).

General conditions for asymptotic normality for dependent sums have been established by Dvoretzky (1972); particular cases that seem useful include mixing CLT's (Iosifescu and Theodorescu, 1969) and methods for associated random variables (Cox and Grimmett, 1984; Newman and Wright, 1982). Once (10) is deemed acceptable, the asymptotic behavior of $\hat{\theta}_J$ is determined by $C_J(\theta)$. When $C_J(\theta)$ is near zero, we can expect the items to behave as though LI were true; when $C_J(\theta)$ is much larger, we should expect item behavior which can be effectively analyzed only with a multidimensional model. We shall return to $C_J(\theta)$ and (9) in Section 7 below.

4 The best local independence model when LI fails

One of the lessons of Section 3 was that it is possible to continue using $d_L = 1$ methods when in fact $d_L > 1$ as long as $d_E = 1$ holds with respect to the trait you want to measure. The construction there makes a particular choice for the "unidimensional IRF's" in a fictitious version of the LI likelihood, namely the "marginal" IRF's in (4) suggested by Stout (1990b). Is this a good choice? Can it be achieved in practice?

It is valuable to set out general forms of the models we are considering. The "ideal" random-effects LI-based model for item response (and other) data in psychological measurement is a mixture model

$$m(\underline{x}_J) = \int r_J(\underline{x}_J \,|\, \theta)\, dF(\theta) \tag{11}$$

where F is the distribution of Θ, and $r_J(\underline{x}_J \,|\, \theta)$ factors as

$$r_J(\underline{x}_J \,|\, \theta) = \prod_{j=1}^{J} r_j(x_j \,|\, \theta)\,. \tag{12}$$

In dichotomous IRT for example, $r_j(x_j \,|\, \theta) = \pi_j(\theta)^{x_j}(1 - \pi_j(\theta))^{1-x_j}$, for some set of ICC's $\pi_j(\theta)$, but the formulation in (11) and (12) works for arbitrary observations $X_1, \ldots, X_J$ on each individual. The main statistical task is inference about each individual's unobserved θ from each individual's observed $\underline{x}_J$, based on the particular form of the right-hand side of (12).

LI models are an attractive and convenient data analysis tool, and are often assumed even though it may be agreed that (12) only approximately fits or reflects the mechanisms underlying the data. Suppose the correct formulation is

$$m(\underline{x}_J) = \int \nu_J(\underline{x}_J \,|\, \theta)\, dF(\theta), \tag{13}$$

where the conditional model for $\underline{X}_J$ given θ is some dependent $\nu_J(\underline{x}_J \,|\, \theta)$ whose structure is not known in detail. How far could an analysis based on (11) and (12) go? It is useful to first know what the "best possible" choice for $r_J(\underline{x}_J \,|\, \theta)$ is. We shall show that the best choice is indeed $q_J(\underline{x}_J \,|\, \theta)$, where

$$q_i(x_i \,|\, \theta) = \int \nu_J(\underline{x}_J \,|\, \theta)\, dx_1 \ldots dx_{i-1}\, dx_{i+1} \ldots dx_n \tag{14}$$

(when the variables are discrete, as in IRT, the multiple integral here is replaced with a multiple sum). Recall the Kullback-Leibler distance

$$D(\nu_J \| r_J) \equiv D_\theta(\nu_J \| r_J) = \int \log \frac{\nu_J(\underline{x}_J \,|\, \theta)}{r_J(\underline{x}_J \,|\, \theta)} \nu_J(\underline{x}_J \,|\, \theta)\, d\underline{x}_J;$$

models that are close in the Kullback-Leibler sense are also close in other more common senses such as mean absolute error.

Proposition 4.1 *$D_\theta(\nu_J \| r_J)$ is minimized over r_J by taking $r_J \equiv q_J$.*

Indeed, following Aitchison (1975), we note that by (14),

$$\begin{aligned} D_\theta(\boldsymbol{\nu}_J \| \boldsymbol{r}_J) &= D_\theta(\boldsymbol{\nu}_J \| \boldsymbol{q}_J) + \sum_{j=1}^{J} \int \log \frac{q_j(x_j \,|\, \theta)}{r_j(x_j \,|\, \theta)} \, q_j(x_j \,|\, \theta) \, dx_j \\ &= D_\theta(\boldsymbol{\nu}_J \| \boldsymbol{q}_J) + \sum_{j=1}^{J} D_\theta(q_j \| r_j), \end{aligned}$$

which is clearly minimized by taking $r_j \equiv q_j$ in each term of the summation at right. This identification is completely general and applies to binary, polytomous, and even continuously-scored items. Further details and extensions of this idea may be found in Section 2 of Clarke and Junker (1991).

In the usual binary IRT context, suppose $d_L = 1$ is violated and let $\boldsymbol{\nu}_J(\underline{x}_J|\theta)$ be the true, locally dependent likelihood for $\underline{x}_J$ given the dominant or target trait θ. Let $\boldsymbol{r}_J(\underline{x}_J|\theta) = \prod_{j=1}^J \pi_j(\theta)^{x_j}(1 - \pi_j(\theta))^{1-x_j}$ be a locally independent likelihood with arbitrary ICC's $\pi_j(\theta)$. The above proposition shows that *if* the desire is to analyze data from $\boldsymbol{\nu}_J(\underline{x}_J|\theta)$ using a "fictional" LI-based likelihood of the form $\boldsymbol{r}_J(\underline{x}_J, \theta)$, *then* the best choices for $\pi_j(\theta)$ in $\boldsymbol{r}_J$ are the true marginal ICC's $P_j(\theta)$ obtained from $\boldsymbol{\nu}_J$ via (14). If θ is the first coordinate of a multidimensional trait vector $\underline{\theta}_1^d$ with respect to which LI holds using multidimensional response functions $\bar{P}_j(\underline{\theta}_1^d)$, then it can be shown that obtaining $P_j(\theta)$ via (14) is equivalent to

$$P_j(\theta) = \int \bar{P}_j(\underline{\theta}_1^d)\omega(\underline{\theta}_2^d|\theta_1 = \theta)\, d\underline{\theta}_2^d.$$

It must be noted that in practice the selection of $\boldsymbol{r}_J$ in (12) is itself often subject to uncertainty, in that the $r_j(\cdot|\theta)$ are typically selected from a parametric family $r_{\underline{\alpha}_j}(\cdot|\theta)$ whose parameters $\underline{\alpha}_1, \ldots, \underline{\alpha}_J$ are estimated from (some subset cf) the data. Tsutakawa and Soltys (1988) and Albert (1991) provide important insights into correctly addressing this issue. On the other hand, there is some evidence that this "best possible" case may approximately be achieved in some large-scale educational testing applications, for example. Wang (1986, 1987) has identified that component ϑ of $\underline{\theta}_1^d = (\theta_1, \ldots, \theta_d)$ in a multidimensional compensatory logistic IRT model which is measured by a fitted unidimensional logistic IRT model. Wang's "reference composite" ϑ is essentially the first component of that rotation of $\underline{\theta}_1^d$ which produces a principal components analysis of the information matrix $I(\underline{\theta}_1^d)$, and she argues that popular IRT model-fitting programs such as LOGIST and BILOG produce stable estimates for item characteristic curves with respect to the reference composite.

5 Structural robustness and posterior distributions

As we have seen, identifying the "product of marginals" $q_J(\underline{x}_J \,|\, \theta)$ as the best LI likelihood to use was easiest to accomplish by considering the general statistical models of Section 4. In the same way, the asymptotic consistency results of Section 3 can best be understood and extended by considering more general models. Let us continue to use the general notation of Section 4.

5.1 Another route to MLE consistency under EI

By analogy with (3), define

$$L_J(\theta) = \log q_J\,(\underline{X}_J\,|\,\theta) = \sum_{j=1}^{J} \log q_j(X_j|\theta); \tag{15}$$

and by analogy with the Kullback-Leibler distance, define

$$D_J(\theta,\tau) \equiv \frac{1}{J}[L_J(\theta) - L_J(\tau)] = \frac{1}{J}\sum_{j=1}^{J} \log \frac{q_j(X_j|\theta)}{q_j(X_j|\tau)}.$$

Let us also abbreviate $q_\theta^J(\cdot) \equiv q_J(\cdot\,|\,\theta)$. $L_J(\theta)$ would be the log-likelihood if LI were true, but we are *not* assuming LI here: i.e., $q_\theta^J(\cdot)$ may not be the true likelihood function. Finally, for each t, define $B_\delta(t) \equiv \{\tau : \ |\tau - t| < \delta\}$. In this general setting we may obtain consistency of the MLE under the following assumptions, *without making explicit assumptions about the violations of LI.*
Assumption C1. For each θ and $t \neq \theta$, there exists $c(t) > 0$, such that

$$\lim_{J\to\infty} P[D_J(\theta,t) > c(t)|\theta] = 1.$$

Assumption C2. For all $t \neq \theta$ and all $\xi > 0$, there exists $\delta > 0$ such that

$$\lim_{J\to\infty} P\left[\inf_{\tau\in B_\delta(t)} D_J(t,\tau) \geq -\xi \middle| \theta\right] = 1.$$

Assumption C3. There exist $c_\Delta > 0$, such that for all $\delta > 0$ and Δ sufficiently large (depending on δ), $\liminf_{J\to\infty} P\left[\inf_{|\tau|>\Delta} D_J(\theta,\tau) > c_\Delta \middle| \theta\right] \geq 1 - \delta$.

Under these assumptions we obtain the following proposition ensures consistency of the MLE. and furthermore gives an "asymptotic convexity" which will be useful later: $L_J(\theta)$ dominates $L_J(\tau)$ as $J \to \infty$, for all τ "away from" θ. The domination will be used in Theorem 5.1 to establish asymptotic normality of the posterior ability distribution constructed from the LI likeihood q_J, even though LI fails. The proof of may be found in Clarke and Junker (1991). Straightforward modifications also give consistency of the posterior mode.

Proposition 5.1 *Under Assumptions C1 through C3, for all* $\epsilon > 0$ *and all* $\delta > 0$, *there exists* $\gamma = \gamma(\epsilon,\delta) > 0$ *such that*

$$\liminf_{J\to\infty} P\left[\inf_{\tau\notin B_\epsilon(\theta)} \frac{1}{J}[L_J(\theta) - L_J(\tau)] \geq \gamma \middle| \theta\right] \geq 1 - \delta \tag{16}$$

and hence the formal MLE $\hat{\theta}_J \xrightarrow{\nu_J} \theta$ *as* $J \to \infty$ *(where "*$\xrightarrow{\nu_J}$*" denotes convergence in* ν_J*-probability).*

Assumptions C1–C3 are what is needed to make the proof work. Ideally we would like Proposition 5.1 under Stout's EI condition, as given in Definition 2.1. An appropriate generalization of EI is the law of large numbers (LLN)

$$\lim_{J\to\infty} \operatorname{Var}\left(\frac{1}{J}\sum_{j=1}^{J} a_j(X_j)\middle|\tau\right) = 0 \tag{17}$$

for all bounded sequences of functions $\{a_j(X_j) : j = 1,2,\ldots\}$ (for any type of X_j's whatsoever). However, the proof of Proposition 5.1 depends on a LLN that holds for sums of log-contrast functions $D_J(\theta,\tau) = (1/J)\sum_{j=1}^{J}\log[q_j(X_j|\theta)/q_j(X_j|\tau)]$, whose summands need not be bounded. (In polytomous or dichotomous IRT settings, since each item has only finitely many possible responses, $D_J(\theta,\tau)$ would have bounded summands, so that (17) suffices; see Proposition 5.2 below.) Lemmas 5.1 and 5.2 show precisely what LLN's are needed in general to obtain Assumptions C1 and C2. The proofs of the lemmas are straightforward bounding in probability arguments which are omitted.

Lemma 5.1 *Suppose*

(a) *For each $t \neq \theta$ there exists $\beta(t) > 0$ such that* $\liminf_{n\to\infty}(1/J)D\left(q_\theta^J \middle\| q_t^J\right) \geq \beta(t)$;

(b) *As $J \to \infty$,* $D_J(\theta,t) - (1/J)D\left(q_\theta^J \middle\| q_t^J\right) \overset{\nu_J}{\to} 0$

Then Assumption C1 holds.

Lemma 5.2 *Suppose that, for all $t \neq \theta$ there exists $\delta_t > 0$ such that*

(a) $\forall\ \xi > 0\ \exists\ \delta \in (0,\delta_t)$, *such that* $\liminf_{J\to\infty}\inf_{\tau\in B_\delta(t)} E[D_J(t,\tau)|\theta] > -\xi$;

(b) $\forall\ \xi > 0\ \exists\ \delta \in (0,\delta_t)$ *such that* $\lim_{J\to\infty} P\left[\sup_{\tau\in B_\delta(t)} |D_J(t,\tau) - E[D_J(t,\tau)|\theta]| < \xi\middle|\tau\right] = 1.$

Then Assumption C2 holds.

Let us specialize these results to a polytomous IRT setting. Recall that each observable variable x_j has k_j values $\xi_{j1},\ldots,\xi_{jk_j}$ (the subject makes one of k_j responses for each item), with each $k_j \leq k_0$ for some fixed $k_0 < \infty$. The LI likelihood is $q_J(\underline{x}_J\,|\,\theta) = \prod_{j=1}^{J} q_j(x_j\,|\,\theta)$, where

$$q_j(x_j\,|\,\theta) = \prod_{l=1}^{k_j} P_{jl}(\theta)^{Y_{jl}},$$

and $Y_{jl} = 1_{\{X_j=\xi_{jl}\}}$.

Proposition 5.2 *Suppose that* EI *and* LAD *hold, and that the response curves P_{jl} satisfy*

$$\text{For each } t,\ 0 < \inf_{j,l} P_{jl}(t) \leq \sup_{j,l} P_{jl}(t) < 1; \tag{18}$$

$$P_{jl}(t) \text{ is continuous at each } t \text{, uniformly in } j \text{ and } l \tag{19}$$

and suppose Assumption C3 holds. Then the "wrong model MLE" $\hat{\theta}_J$ is ν_J-consistent for θ, as $J \to \infty$.

Proof. We will verify the conditions of Lemma 5.1 and Lemma 5.2. It follows from an inequality of Csiszar (1975), $D\,(f\|g) \ge (1/4)\,[\int |f(t)-g(t)|\,dt]^2$,

$$\begin{aligned}\frac{1}{J} D\left(q_\theta^J \middle\| q_\tau^J\right) &= \frac{1}{J}\sum_{j=1}^{J} D\,(q_{j,\theta}\| q_{j,\tau})\\ &\ge \frac{1}{4}\frac{1}{J}\sum_{j=1}^{J}\left[\sum_{l=1}^{k_j} |P_{jl}(\theta)-P_{jl}(\tau)|\right]^2,\end{aligned} \qquad (20)$$

which is bounded away from zero under LAD (consider $\{a_{jl}\}$ for which $a_{jk_j} = 1$, and $a_{jl} \equiv 0$ for all $l < k_j$). This is (a) of Lemma 5.1. On the other hand, (b) of Lemma 5.1 follows from Definition 2.1 and (18), since the summands of $D_J(\theta,\tau)$ are bounded.

The continuity condition (a) of Lemma 5.2 follows from (19). (b) of Lemma 5.2 requires that

$$\lim_{J\to\infty} P\left[\sup_{\tau\in B_\delta(t)} |D_J(t,\tau) - E[D_J(t,\tau)|\theta]| < \epsilon \,\middle|\, \theta\right] = 1$$

for every ϵ and appropriate δ. The expression in absolute values may be written as

$$\frac{1}{J}\sum_{j=1}^{J}\sum_{l=1}^{k_j}[Y_{jl} - P_{jl}(\theta)]\log\frac{P_{jl}(t)}{P_{jl}(\tau)}$$

which will tend to zero uniformly in $\tau \in B_\delta(t)$ by Definition 2.1, (18) and (19). □

Assumption C3 may often be verified directly. Consider the case of binary response data, in which $k_j \equiv 2$, $\xi_{j1} \equiv 0$ and $\xi_{j2} \equiv 1$, and the response curves are of the three parameter logistic form

$$P_j(\theta) = c_j + (1-c_j)\frac{1}{1+\exp\{-a_j(\theta-b_j)\}}.$$

Then $D_J(\theta,\tau) = (1/J)\sum_1^J t_j(\theta) - \iota_j(\tau)$, where

$$t_j(\tau) = X_j \log\frac{c_j + e^{a_j(\tau-b_j)}}{1-c_j} - \log\left[1 + \frac{c_j+e^{a_j(\tau-b_j)}}{1-c_j}\right].$$

Hence

$$\begin{aligned}\lim_{\tau\to\infty} -t_j(\tau) &= \begin{cases} 0, \text{ if } X_j = 1,\\ \infty, \text{ if } X_j = 0;\end{cases}\\ \lim_{\tau\to-\infty} -t_j(\tau) &= -\log c_j^{X_j}(1-c_j)^{1-X_j},\end{aligned}$$

and we see that Assumption C3 holds as long as $P[X_j = 1\,\forall j|\theta] = P[X_j = 0\,\forall j|\theta] = 0$; this in turn follows from Definition 2.1 and (18), which merely requires that the a_j's b_j's and c_j's do not "wander off" to the edges of their parameter spaces.

5.2 Asymptotic posterior normality under EI

We now turn to the possibility of basing inference for θ on the formal posterior distribution

$$\omega_q(\theta \mid \underline{x}_J) = \frac{q_J(\underline{x}_J \mid \theta)\omega(\theta)}{\int_{-\infty}^{\infty} q_J(\underline{x}_J \mid \tau)\omega(\tau)\, d\tau}, \tag{21}$$

where $\omega(\theta)$ is the prior density on θ. Of course, the true posterior distribution is

$$\omega_\nu(\theta \mid \underline{x}_J) = \frac{\nu_J(\underline{x}_J \mid \theta)\omega(\theta)}{\int_{-\infty}^{\infty} \nu_J(\underline{x}_J \mid \tau)\omega(\tau)\, d\tau}.$$

The point once again is to see whether a "wrong model ana'vsis" based on the LI likelihood q_J can work when ν_J is the correct conditional law. Let us make the following regularity assumptions.

Assumption PN1. Let $I_j(\theta) = E\left[(\partial \log q_j(X_j|\theta)/\partial\theta)^2 \middle| \theta\right]$ and $\overline{I}_J(\theta) = (1/J)\sum_1^J I_j(\theta)$. We assume there exist $0 < \epsilon_\theta \leq M_\theta < \infty$ such that $\epsilon_\theta \leq \overline{I}_J(\theta) \leq M_\theta$, for all large n.

Assumption PN2. $\int \partial^2 q_j(x|\theta)/\partial\theta^2 dx = 0$.

Assumption PN3. $M_{\epsilon,j}(x,\theta) = \sup_{\tau \in B_\epsilon(\theta)} |\partial^2 \log q_j(x|\tau)/\partial\tau^2 - \partial^2 \log q_j(x|\theta)/\partial\theta^2|$ is bounded uniformly in x and j, for small $\epsilon > 0$, ; and for $\overline{M}_J(\epsilon,\theta) = (1/J)\sum_1^J M_{\epsilon,j}(X_j,\theta)$,

$$\lim_{\epsilon \to 0} \limsup_{J \to \infty} E\left[\overline{M}_J(\epsilon,\theta) \middle| \theta\right] = 0.$$

Assumption PN4. The prior density $\omega(\tau)$ is positive and continuous throughout a small neighborhood of θ.

Theorem 5.1 *Assume EI as in (17), and the conclusion of Proposition 5.1. Under the additional assumptions PN1 through PN4, for all $a < b$,*

$$\int_{\hat{\theta}_J + a\sigma_J}^{\hat{\theta}_J + b\sigma_J} \omega_q(\theta \mid \underline{X}_J)\, d\theta \overset{\nu_J}{\to} \Phi(b) - \Phi(a) \tag{22}$$

as $J \to \infty$, where $\sigma_J = \{-L_J''(\hat{\theta}_J)\}^{-1/2}$, and $\Phi(\cdot)$ is the the standard normal c.d.f.

Hence, in contrast to Theorem 3.1, which shows that the asymptotic distribution of the MLE is sensitive to departures from strict unidimensionality, Theorem 5.1 suggests that the asymptotic posterior ability distribution cannot "detect" such departures. While this may initially seem to be good news, it actually undermines the desirability of basing inference about θ on a wrong-model posterior. We shall return to this point at the end of the section.

The proof of this result, and extensions to situations in which EI fails, may be found in Clarke and Junker (1991). Chung (1991) has independently produced a proof of this result in the traditional, LI-based, dichotomous IRT setting. In both cases, the calculations are modeled after Walker (1969). Straightforward modifications give consistency of the posterior mean and higher posterior moments.

The next proposition specializes the result to essentially unidimensional polytomous IRT models.

Proposition 5.3 *Suppose, in addition to the assumptions of Proposition 5.2, that*

$$\frac{\partial^2}{\partial\theta^2} \log P_{jl}(\theta) \text{ is bounded pointwise in } \theta, \text{ uniformly in } j \text{ and } l. \tag{23}$$

Then, in the sense of (22),

$$\mathcal{L}_q \left\{ \frac{\Theta - \hat{\theta}_J}{\sigma_J} \middle| \underline{x}_J \right\} \xrightarrow{\nu_J} N(0,1).$$

Proof. Assumptions PN2 and PN4 are usually true "by fiat," so only it is only interesting to consider Assumption PN1 and Assumption PN3. Proposition 4.1 of Junker (1991b) shows that Assumption PN1 holds under LAD and differentiability conditions (the argument is similar to the one bounding (20) away from zero). The uniform continuity condition of Assumption PN3 focuses on a locally uniform bound for

$$\left| \sum_{l=1}^{k_j} Y_{jl} \left[\frac{\partial^2}{\partial\tau^2} \log P_{jl}(\tau) - \frac{\partial^2}{\partial\theta^2} \log P_{jl}(\theta) \right] \right|; \tag{24}$$

which follows from (23), due to the boundedness of the Y_{jl}'s. □

Example 5.1 Stout (1990b) and Junker (1991b) consider binary responses $X_1, X_2, X_3, \ldots$, having the same response curve $P[X_j = 1 \mid \theta] \equiv \theta$. Suppose that the items are arranged in successive groups of g_o items as $X_1, X_2, \ldots, X_{g_o}$; $X_{g_o+1}, X_{g_o+2}, \ldots, X_{2g_o}$; etc., such that different groups of g_o items are independent of one another, given θ, and items within a single group are positively correlated, given θ, and with

$$\mathrm{Corr}(X_i, X_j|\theta) = \begin{cases} c \text{ if } X_i \text{ and } X_j \text{ are in the same group,} \\ 0 \text{ if not,} \end{cases}$$

for some fixed $c \in (0,1]$. This ν_J is a naive model for a paragraph comprehension test in which several paragraphs are presented and g_o questions are asked for each paragraph. Here, θ represents a trait common to all the items, which we might wish to think of as reading comprehension; and the nonzero correlations are induced by nuisance traits, for example, specific knowledge about the subject matter of the paragraph at hand.

EI and LAD hold in this case, and it follows from Proposition 5.2 and Theorem 3.3 that

$$\sqrt{J}(\hat{\theta}_J - \theta) \xrightarrow{\mathcal{L}} N(0, \sigma^2),$$

where $\hat{\theta}_J = \bar{x}_J$, and $\sigma^2 \equiv \theta(1-\theta)[1+c(g_0-1)]$ is somewhat inflated over the anticipated asymptotic variance $\theta(1-\theta)$ under LI. On the other hand, it follows from Proposition 5.3 that

$$\mathcal{L}\left\{ \sqrt{J} \frac{\Theta - \hat{\theta}_J}{\sqrt{\hat{\theta}_J(1-\hat{\theta}_J)}} \underline{x}_J \right\} \xrightarrow{P} N(0,1). \qquad \square$$

In Example 5.1, the asymptotic distribution of the MLE has an inflated variance, due to the departure from strict unidimensionality, but the asymptotic posterior does not. Moreover, careful examination of Theorem 3.1 and Theorem 5.1 makes it clear that the asymptotic distribution of the MLE is always potentially sensitive to any "local dependence" in the data, even when $d_E = 1$ (Definition 2.2) holds, while the asymptotic posterior distribution under $d_E = 1$ never is. Clarke and Junker (1991) also examine this phenomenon in some $d_E > 1$ situations. It is widely believed that the two paradigms, likelihood-based inference and posterior-based inference, are philosophically different but "asymptotically the same", except in bizarre situations. But the perfectly reasonable desire to analyze IRT data using unidimensional models that are tolerant of minor violations of strict unidimensionality has lead us into a situation in which the asymptotics come out differently, even for "typical" cases. How can we make sense of this?

On the one hand, the LI-based MLE $\hat{\theta}_J$ is really an M-estimator with a particular choice of objective function, namely the product of the one-dimensional data marginals of ν_J, which we have denoted q_J. Thus we may interpret the asymptotic distribution of the M-estimator $\hat{\theta}_J$ as a measure of estimation error under ν_J without difficulty; in particular we need not assume that the data actually came from q_J to arrive at this interpretation.

On the other hand, our approximation to the LI-based posterior shows that it concentrates at the LI-based M-estimator—cf. equation (22)—but its "asymptotic rate of concentration" is harder to interpret: LI-based asymptotic posterior standard errors say how much the LI-based posterior is concentrated around the M-estimator, but not how much the LI-based posterior is concentrated around the θ which "generated" $\underline{x}_J$. If an LI model really held, then Bayes' rule would allow us to interpret the LI-based posterior, and hence its asymptotically normal approximation, in the usual sense of updating belief about where θ was after looking at the data. If LI does not hold, then we cannot appeal to Bayes' rule for this interpretation, and the LI-based posterior is interesting only because it corresponds to what is done in practice. Perhaps the only justifiable interpretation of ω_q is a counterfactual: "If LI *were* true, this is where we would think θ was."

Although both MLE and Bayes paradigms lead to consistent estimators when the LI-based likelihood q_J is substituted for the true dependent likelihood ν_J, correct calculation and interpretation of the variability of the estimators depends on a more careful analysis of the stochastic behavior of the data-generating mechanism. Detecting situations in which this must be done is the major goal of the work reported in Sections Section 6 and Section 7.

6 A global index of unidimensionality

Stout (1987) proposes a statistical test of unidimensionality for binary IRT data, which has been further investigated by Stout and Nandakumar (1987, 1989, 1991a, 1991b). The test statistic is based on a quantity which may be interpreted as an estimate of the measure

$$\epsilon_J = \int \binom{J}{2}^{-1} \sum\sum_{1 \leq i < j \leq J} |\mathrm{Cov}\,(X_i, X_j|\theta)| f(\theta) d\theta$$

of unidimensionality of IRT data. Note that under $d_L = 1$ the covariances are identically zero, so that $\epsilon_J \equiv 0$. Under $d_E = 1$, the covariances tend to zero as J grows, and hence $\epsilon_J \approx 0$ for $d_E = 1$

data. If the data is dramatically multidimensional, the covariance will be prediminantly nonzero and we expect $\epsilon_J \gg 0$. This measure can be estimated directly with the index

$$\hat{\epsilon}_J^0 = \frac{1}{N}\sum_{k=0}^{J} N_k \binom{J}{2}^{-1} \sum\sum_{1\le i<j\le\ell} |\widehat{\text{Cov}}\ (X_i, X_j|X_+ = k)|$$

where X_+ is the total score on the whole test, N_k is the number of examinees with total score k out of J on the whole test; and the estimate $\widehat{\text{Cov}}\ (X_i, X_j|X_+ = k)$ is obtained in the usual way as $(1/N_k)\sum_{n=1}^{N_k}(x_{ni} - \bar{x}_i)(x_{nj} - \bar{x}_j)$, with $\bar{x}_j = (1/N_k)\sum_{n=1}^{N_k} x_{nj}$ (the sums extending only over examinees in the k^{th} cohort).

The ideal behavior of this index should be

$$\begin{aligned} \hat{\epsilon}_J^0 &\approx 0 \text{ if } d_E = 1; \\ \hat{\epsilon}_J^0 &\gg 0 \text{ if } d_E > 1. \end{aligned}$$

Initial study of this index showed that $\hat{\epsilon}_J^0$ was greatly inflated in unidimensional cases. The inflation could be attributed to either of two causes: some covariances were nonzero because of natural random variability in the data; and others were nonzero because, in many strictly unidimensional models, Cov $(X_i, X_j|X_+ = k) < 0$ may occur *even though* Cov $(X_i, X_j|\Theta = \theta) \equiv 0\ \forall\ \theta$ (see Junker (1991a) for a theoretical discussion of this point). Since the absolute values of the covariances are summed in calculating the index $\hat{\epsilon}_J^0$, these latter negative covariances, which are in fact *due to* unidimensionality, were counted *against* unidimensionality in the index.

To remedy the situation, the following four-step construction was formulated:

1. Perform a principal components factor analysis of the tetrachoric correlation matrix and retain the list of second factor loadings, $\{\lambda_{j2} : j = 1, \ldots, J\}$.

2. Cast out individual items X_j for which $|\lambda_{j2}| < M$ for some fixed cutoff M.

3. For each $k = 0, \ldots, J$, obtain covariance estimates $\widehat{\text{Cov}}\ (X_i, X_j|X_+ = k)$ for all the item pairs left after applying Step 2. (Note: X_+ is formed from *all* the items, but only covariances among the items remaining after Step 2 are calculated in each X_+ cohort.)

   (a) If $\lambda_{i2}\cdot\lambda_{j2}$ has the same sign as the estimate $\widehat{\text{Cov}}\ (X_i, X_j|X_+ = k)$, retain this covariance; otherwise cast it out.

   (b) Calculate

   $$\hat{\epsilon}_J(k) = \binom{J}{2}^{-1} \sum\sum_{\text{remaining pairs}} |\widehat{\text{Cov}}\ (X_i, X_j|X_+ = k)|$$

   where the sum is over all those pair remaining after Steps 2 and 3a.

4. Calculate the new index

$$\hat{\epsilon}_J = \frac{1}{N}\sum_{k=0}^{J} N_k \hat{\epsilon}_J(k).$$

A rationale for this construction is most easily seen by contrasting a strictly unidimensional test with a test consisting of two strictly unidimensional subtests, say half "math" items and half "verbal" items. In the strictly unidimensional case, the first factor of a principal-components factor analysis of the tetrachoric correlations will be close to the true ability factor underlying the test, and the second factor will pick up only random variation in the data. Thus many of the second factor loadings λ_{j2} should be quite small; these items are automatically dropped from the analysis by Step 2. Many pairs of the remaining items will have $\widehat{\mathrm{Cov}}\ (X_i, X_j|X_+ = k) < 0$, and approximately half of these should be dropped because the "random" second factor loadings should satisfy $\lambda_{i2} \cdot \lambda_{j2} > 0$ about half the time. Thus most of the covariances are not included in the calculation in Steps 3b and 4, and therefore $\hat{\epsilon}_J \approx 0$ in the unidimensional case.

In the case of two different, strictly unidimensional subtests, the first factor of a principal-components factor analysis of the tetrachoric correlations will be a general factor correlating highly with the number-right score. The second factor will be a "contrast" (or bipolar) factor for which items in one subtest, say the "math" items, will load positively; and items in the other subtest, say "verbal" items, will load negatively. A few items will be cast out in Step 2 again because they do not load heavily enough on the contrast factor. Of those remaining, consider separately the cases $\lambda_{i2} \cdot \lambda_{j2} > 0$ and $\lambda_{i2} \cdot \lambda_{j2} < 0$. If the product is positive, both items probably come from the same subtest and we expect $\widehat{\mathrm{Cov}}\ (X_i, X_j|X_+ = k) > 0$ (since X_+ is summed over both subtests it is measuring "$\theta_{\text{math}} + \theta_{\text{verbal}}$"; if the items are both "math", the "verbal" component of X_+ will tend to make the covariance positive, and vice-versa). We would like to keep this covariance in the calculation for $\hat{\epsilon}_J$ and this is what Step 3a does. On the other hand, if the product is negative, the items probably come from different subtests and we expect $\widehat{\mathrm{Cov}}\ (X_i, X_j|X_+ = k) < 0$ (this is the non-unidimensional behavior that tests of "conditional association", Holland and Rosenbaum, 1986, are designed to detect). We would also like to keep this covariance in the sum, and Step 3a does this for us too. Thus most of the covariances are included in the calculation in Steps 3b and 4, and therefore $\hat{\epsilon}_J \gg 0$ in the non-unidimensional case.

Preliminary simulation and real-data studies with the index $\hat{\epsilon}_J$ are quite promising, as Tables 1 and 2 show. In Table 1, the first simulation marked "$d = 1$" is based upon a two parameter logistic model with discriminations $a_j \sim N(1.28, (0.8)^2)$, sampled until $0.5 \leq a_j \leq 3$; and difficulties $b_j \sim N(-0.12, (0.84)^2)$, sampled until $-3 \leq b_j \leq 3$. The simulations marked "$d = 2$" are based on tests consisting of two pure subtests with correlation $\rho_{\theta_1, \theta_2} = 0.3$ between traits, and item parameters generated according to the same distributions as in the $d = 1$ case, except as noted. The simulations marked ASVAB AR and ASVAB AS are generated according to the three parameter logistic model, using the fixed item parameter estimates for particular administrations of the Armed Services Vocational Aptitude Battery, Arithmetic Reasoning and Auto Shop sections, attributed to Bock by Nandakumar (1987).

The most striking aspect of Table 1 is the marked contrast in the values of $\hat{\epsilon}_J$ between the one- and two-dimensional cases. This certainly supports the rationale behind the construction of $\hat{\epsilon}_J$ above. It is also interesting to note the progression of values of the index as the second factor loading cutoff value M increases from 0.0 to 0.2. Clearly, in this range, increasing M improves the performance of $\hat{\epsilon}_J$ in the unidimensional case without degrading its performance on strongly two-dimensional data. By increasing M to 0.2, we are able to effectively decrease the propensity

Simulated data sets		M:	.00	.10	.15	.20
	J	N	$100\hat{\epsilon}_J$			
$d = 1$	40	2000	.84	.21	.15	.10
$d = 2$, $\sigma_a = 0.8$	20+20	2000	2.29	2.29	2.22	2.14
$d = 2$, $\sigma_a = 0.6$	20+20	2000	2.68	2.68	2.68	2.68
$d = 2$, $\sigma_a = 0.4$	20+20	2000	2.38	2.38	2.38	2.38
ASVAB AR ($d = 1$)	30	2000	.72	.49	.20	.06
ASVAB AS ($d = 1$)	25	2000	.75	.16	.07	.05

Table 1: $\hat{\epsilon}_J$, applied to simulated data sets.

Real data sets		M:	.00	.10	.15	.20
	J	N	$100\hat{\epsilon}_J$			
ACT F29B (math)	40	2491	.94	.55	.25	.07
ACT F29C (math)	40	2494	.96	.52	.26	.10
AR 10 (ASVAB)	30	1984	.74	.28	.16	.04
AR 12 (ASVAB)	30	1961	.74	.23	.17	.11

Table 2: $\hat{\epsilon}_J$, applied to real data sets.

for making a Type I error without noticably affecting Type II error.

The $\hat{\epsilon}_J$ index has also been applied to some real data sets, with the results in Table 2. The first two lines of the table are from the Mathematics section of the ACT (American College Testing) Assessment, Forms 29B and 29C. The next two lines are Arithmetic Reasoning sections of the ASVAB.

These preliminary results show that $\hat{\epsilon}_J$ is a promising global index of unidimensionality. Clearly there is much more work to be done in understanding the performance of the index through simulation experiments and in applying the index to real data sets. It would also be interesting to compare $\hat{\epsilon}_J$ to the $Q3$ measure of LI model fit developed by Yen (1984). A more finely-tunable version of $\hat{\epsilon}_J$, in which the "cutoff" parameter M may take different values depending on the signs of λ_{i2} and λ_{j2}, will also be explored in future work.

7 A local index of unidimensionality

An alternative to developing a single global index of unidimensionalty is to try to develop an index or diagnostic criterion which helps us understand the nature of violations of strict unidimensionality, or identifies areas of the "unidimensional" ability scale in which ability estimation based on strictly unidimensional assumptions may not succeed. The index $C_J(\theta)$ as described in (9) is such an index.

Under strict unidimensionality the asymptotic standard error of the MLE, for example, is

$$\begin{aligned} SE(\hat{\theta}_J) &= \sqrt{J\text{Var}\,(\hat{\theta}_J|\theta)} \\ &= \sqrt{1/\bar{I}_J(\theta)}, \end{aligned}$$

where θ is the true ability value for the examinee "generating" the response sequence from which $\hat{\theta}_J$ is calculated. However we saw in Section 3 that when strict unidimensionality fails, a correction using $C_J(\theta)$ from (9) is required:

$$SE^*(\hat{\theta}_J) = \sqrt{\frac{\bar{I}_J(\theta) + C_J(\theta)}{\bar{I}_J(\theta)^2}}.$$

Another way to measure the change in accuracy of ability estimation is to consider a corrected information function

$$\bar{I}^*_J(\theta) = \frac{\bar{I}_J(\theta)^2}{\bar{I}_J(\theta) + C_J(\theta)}.$$

Thus $C_J(\theta)$, if it could be estimated, would help us to interpret exactly when ability estimation based on a unidimensional model behaves as though the data were strictly unidimensional. Indeed there are three interesting cases:

I. When $d_L = 1$ holds exactly, $C_J(\theta) \equiv 0$ for all θ, and the "corrected" standard error and information functions reduce to the familiar traditional forms. More generally if $C_J(\theta)$ hovers near zero over the range of values of θ of interest, then it would seem reasonable to pursue ability estimation assuming that the data strictly satisfies $d_L = 1$.

II. If $C_J(\theta)$ is clearly distinct from zero, but not large for most values of θ of interest, it may be desirable to continue to use unidimensional ability estimation methods, but use the corrected standard error SE^* in assessing the accuracy of ability estimation.

III. If $C_J(\theta)$ is quite large for many values of θ of interest, it is probably most desirable to abandon unidimensional modeling completely and develop a multidimensional model for the data set.

In order to estimate $C_J(\theta)$ and SE^*, the following three quantities must be estimated (see (9) on p. 6):

1. Item characteristic curves $P_j(\theta)$;

2. Derivatives of item log-odds-ratios $\lambda'_j(\theta) = P'_j(\theta)/(P_j(\theta)(1 - P_j(\theta))$;

3. "Local" item covariances Cov $(X_i, X_j|\theta)$.

Estimates of the average test information $\bar{I}_J(\theta)$, the usual asymptotic MLE standard error SE, $C_J(\theta)$ itself, and the corrected standard error SE^* may be obtained as straightforward combinations of the above quantities.

In general there are two ways to tackle this problem. One is to explicitly model for the anticipated dependence in the data. This is the approach of Gibbons, Bock and Hedeker (1989), for example. Consider a multidimensional compensatory IRT model with normal ogive item characteristic curves (ICC's). An appropriate—and equivalent—reformulation of the problem is to consider underlying "propensity variables" $Y_1, Y_2, \ldots, Y_J$, such that

$$X_j = 1 \text{ if and only if } Y_j > \gamma_j$$

where, $Y_1, Y_2, \ldots, Y_J$ are independent, $N(\sum_1^d \lambda_{jm}\theta_m, 1 - \sum \lambda_{jm}^2)$ random variables, given the multidimensional latent trait $\underset{\sim}{\theta}_1^d = (\theta_1, \ldots, \theta_d)$; and $\underset{\sim}{\theta}_1^d \sim N(\underset{\sim}{0}, I_{d\times d})$. The thresholds γ_j correspond to difficulty parameters, and the coefficients λ_{jm} correspond to discrimination parameters. (This is also the formulation of item factor analysis which underlies the factor analysis of tetrachoric correlations in Section 6 above, in which the λ_{jm} are the m^{th} factor loadings). Gibbons, Bock and Hedeker (1989) consider a slightly different formulation of the problem, in which $\theta \sim N(0,1)$ is unidimensional, and $(Y_1, \ldots, Y_J) \sim N((\lambda_1\theta, \ldots, \lambda_J\theta), \Sigma)$, given θ, for some covariance matrix Σ. Clearly, if the λ_j, γ_j and Σ could be estimated, estimates of $C_J(\theta)$, SE and SE^* would follow naturally from these and the known normal ogive form of the ICC's. However in our early attempts to use this model, we have found the parameter estimates to be too unstable, especially for tests of more than a handful of items, to be of use. Nevertheless this is an interesting and attractive approach which ought to receive more attention in the future.

A second approach to the problem of estimating ICC's and local item correlations for possibly non-unidimensional data may be based on the nonparametric rank regression methods of Ramsay (1990). Two important observations underlie Ramsay's approach. The first is that we can sidestep the usual identifiability problem for the ability distribution—one aspect of which is that ability estimates are only determined up to rank ordering in the usual IRT formulations—by fixing the distribution of ability (estimates) in advance and allowing quite general ICC shapes in order fit the observed item response distribution. Ramsay's second observation is that very simple ability estimates, based on number-right scores and similar quantities, are quite adequate as "initial guesses" for constructing ICC estimates. This second observation harmonizes nicely with the observation of Stout (1990) that, under essential unidimensionality, $\overline{P}_J^{-1}(\overline{X}_J) \xrightarrow{p} \theta$, as well as with the more traditional view that when an unrotated principal-components factor analysis of binary items is performed, the first factor (corresponding to the largest eigenvalue) is usually strongly related to the total test score on the test (whether or not the test is unidimensional).

In our implementation of Ramsay's method, we obtained approximately $N(0,1)$-distributed ability estimates by inverse-probability transforms of the ranks of examinees' number-right scores. Let us call these crude ability estimates $t_1, t_2, \ldots, t_N$. Also, let $w(t)$ be the standard normal density. Then $P_j(\theta)$ can be estimated nonparametrically using the Nardaraya-Watson kernel regression formula

$$\hat{P}_j(\theta) = \frac{\sum_{n=1}^N x_{nj} w((t_n - \theta)/h)}{\sum_{n=1}^N w((t_n - \theta)/h)},$$

where $h > 0$ is a "window width" or "bandwidth" tuning parameter, and $(x_{n1}, x_{n2}, \ldots, x_{nJ})$ is the observed response pattern of the n^{th} examinee, $n = 1, \ldots, N$. The derivatives $P_j'(\theta)$ may be crudely estimated by considering equally-spaced points $s_1, \ldots, s_K$ in the interval $[-3,3]$ and calculating the difference quotients

$$\hat{P}_j'(s_k) = \frac{\hat{P}_j(s_{k+1}) - \hat{P}_j(s_k)}{s_{k+1} - s_k}.$$

(More sophisticated kernel estimates of the derivatives can be obtained, but these crude estimates were quick and adequate for our initial investigations.) Finally, Ramsay's method was extended to calculate the local item covariances according to the formula

$$\widehat{\text{Cov}}\ (X_i, X_j|t) = \frac{\sum_{n=1}^{N} x_{ni}x_{nj}w((t_n - \theta)/h)}{\sum_{n=1}^{N} w((t_n - \theta)/h)} - \hat{P}_i(t)\hat{P}_j(t).$$

In our work, all quantities were evaluated at $K = 32$ equally-spaced points $s_1, s_2, \ldots, s_{32}$ in [-3,3], with window-width $h = 0.3$. Calculations were performed in the statistical package "New S" on DECstation 3100's. With default memory allocations in S, data sets with up to $N = 500$ examinees and up to approximately $J = 50$ items could be examined. Work on $C_J(\theta)$ is still in preliminary stages, but we provide some illustrative examples.

To illustrate the method, let us simulate one- and two-dimensional tests with $J = 32$ items and $N = 500$ examinees, with compensatory two parameter logistic item parameters as in Table 3 (examinee abilities in all dimensions are sampled from $N(0,1)$ as usual). Note that the one-dimensional item parameters are the average of the two-dimensional parameters.

Since this work is in part a replication of Ramsay's method it is interesting to see how well the rank regression method recovers ICC's. In Figure 1 we have graphed a few one dimensional logistic ICC's (symbol ".") using the parameters on the left in Table 3. Overlaid on these are the unidimensional rank regression ICC estimates from $N = 500$ simulated examinees taking the one dimensional items in Table 3 (symbol "*"). It seems that the rank regression ICC estimates recover the original one dimensional ICC's quite well.

On the other hand, consider Figure 2. The ICC's marked "." are the marginal ICC's $P_j(\theta_1) = \int P_j(\theta_1, \theta_2)\omega(\theta_2|\theta_1)d\theta_2$, where $P_j(\theta_1, \theta_2)$ are compensatory logistic ICC's using the item parameters on the right in Table 3. Overlaid (symbol "*") are the unidimensional rank regression ICC estimates from Ramsay's method (again using $N = 500$ simulated examinees). As expected, the estimated ICC's in Figure 2 do not match the theoretical ICC's nearly as well as in Figure 1. (This assumes that θ_1 is the ability we intend to measure; in the future we would prefer to compare the rank regression ICC estimates with marginal ICC's for Wang's (1986, 1987) "reference composite").

To illustrate the summands for our estimate of $C_J(\theta)$ we may consider Figures 3 and 4, in which rank-regression estimates of the covariances Cov $(X_i, X_j|\theta)$ (symbol ".") and the "weighted" covariances $\lambda_i'(\theta)\lambda_j'(\theta)$Cov $(X_i, X_j|\theta)$ (symbol "*") are depicted. Since the data for Figure 3 comes from a strictly unidimensional model, we know that the theoretical value of Cov $(X_i, X_j|\theta)$ is zero in Figure 3 (which is shown as a horizontal line). The estimated covariances do indeed hover around zero (note that the vertical scale typically ranges from about -0.04 to $+0.10$).

On the other hand, we expect Cov $(X_i, X_j|\theta)$ to be positive in Figure 4, because the data comes from a two dimensional model and we are only conditioning on a one-dimensional θ. The estimated covariances in Figure 4 do seem to range about twice as far from zero, on average, as the covariance estimates for unidimensional data did. The fact that the estimates sometimes dip below zero in

One dimensional data parameters

j	a_j	b_j	c_j
1	1.54	0.453	0
2	0.75	0.404	0
3	0.78	0.038	0
4	1.49	−0.185	0
5	1.06	−0.331	0
6	0.96	0.023	0
7	0.99	−0.093	0
8	1.20	−1.104	0
9	1.38	0.300	0
10	1.28	0.106	0
11	1.43	0.327	0
12	1.47	−0.322	0
13	1.20	−0.339	0
14	1.80	−0.355	0
15	1.87	−0.174	0
16	1.38	−0.212	0
17	1.67	−0.089	0
18	1.43	−0.499	0
19	1.45	−0.114	0
20	1.01	−0.094	0
21	1.10	−0.097	0
22	1.85	−0.365	0
23	1.48	−1.062	0
24	0.74	0.015	0
25	1.46	−0.250	0
26	1.54	−0.542	0
27	1.78	−0.265	0
28	1.35	−0.573	0
29	1.11	0.013	0
30	1.52	0.537	0
31	1.00	−0.082	0
32	1.00	−0.163	0

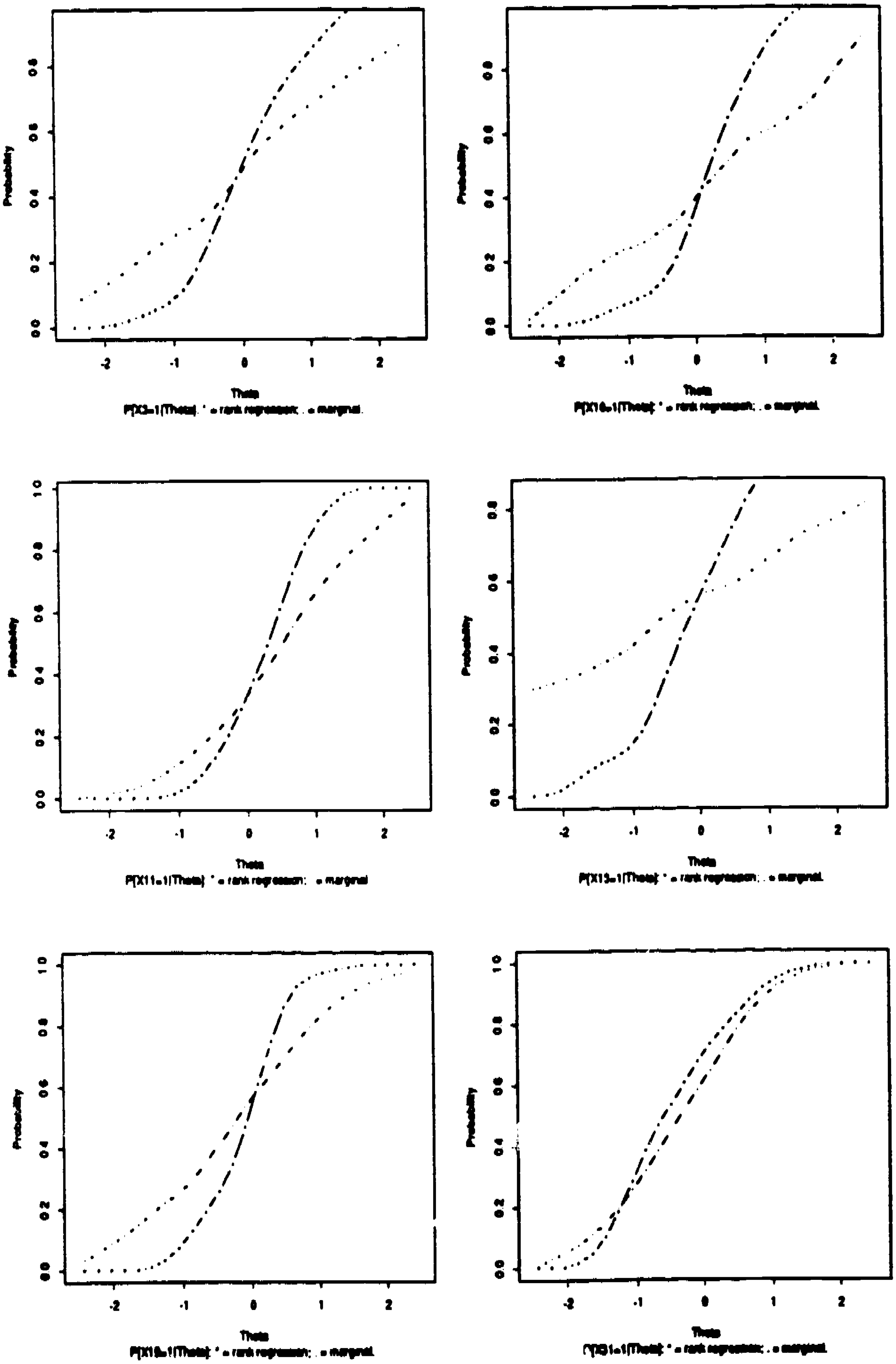
$J = 32$, $N = 500$, $d = 1$

Two dimensional data parameters

j	a_j		b_j		c_j
1	1.54	1.54	−0.0201	0.9254	0
2	0.58	0.91	0.5282	0.2794	0
3	0.70	0.87	−0.0868	0.1632	0
4	1.45	1.53	−0.0400	−0.3299	0
5	1.02	1.09	0.5738	−1.2352	0
6	1.02	0.90	−0.0441	0.0903	0
7	0.99	0.99	−0.8639	0.6787	0
8	1.87	0.53	−1.3073	−0.9015	0
9	1.58	1.19	0.8938	−0.2934	0
10	1.02	1.55	−0.1484	0.3596	0
11	1.53	1.34	−0.3824	1.0370	0
12	0.81	2.14	−0.7693	0.1253	0
13	0.62	1.78	−0.8339	0.1558	0
14	1.87	1.74	−0.3210	−0.3886	0
15	1.59	2.14	−0.1208	−0.2263	0
16	1.75	1.02	0.0816	−0.5054	0
17	1.64	1.71	−0.3706	0.1931	0
18	1.93	0.93	−1.3920	0.3943	0
19	1.44	1.46	0.4418	−0.6704	0
20	0.87	1.15	0.0551	−0.2432	0
21	1.05	1.14	0.4246	−0.6180	0
22	2.03	1.67	0.6099	−1.3403	0
23	1.88	1.09	−1.3022	−0.8211	0
24	0.82	0.65	0.6171	−0.5866	0
25	1.87	1.05	−0.0511	−0.4483	0
26	1.71	1.37	−0.8606	−0.2236	0
27	2.14	1.42	−0.1327	−0.3973	0
28	1.26	1.43	−0.7932	−0.3521	0
29	1.29	0.93	−0.0478	0.0741	0
30	1.78	1.26	0.5742	0.5003	0
31	1.32	0.69	−0.6269	0.4628	0
32	0.65	1.35	−0.3338	0.0078	0

$J = 32$, $N = 500$, $d = 2$, $\rho = 0$

Table 3: Item parameters for illustration.



P[X3=1|Theta]. * = rank regression; . = marginal.

P[X16=1|Theta]: * = rank regression; . = marginal.

P[X11=1|Theta]: * = rank regression; . = marginal

P[X13=1|Theta]: * = rank regression; . = marginal.

P[X19=1|Theta]: * = rank regression; . = marginal.

P[X31=1|Theta]: * = rank regression; . = marginal.

Figure 2: One-dimensional ICC's for two-dimensional data.

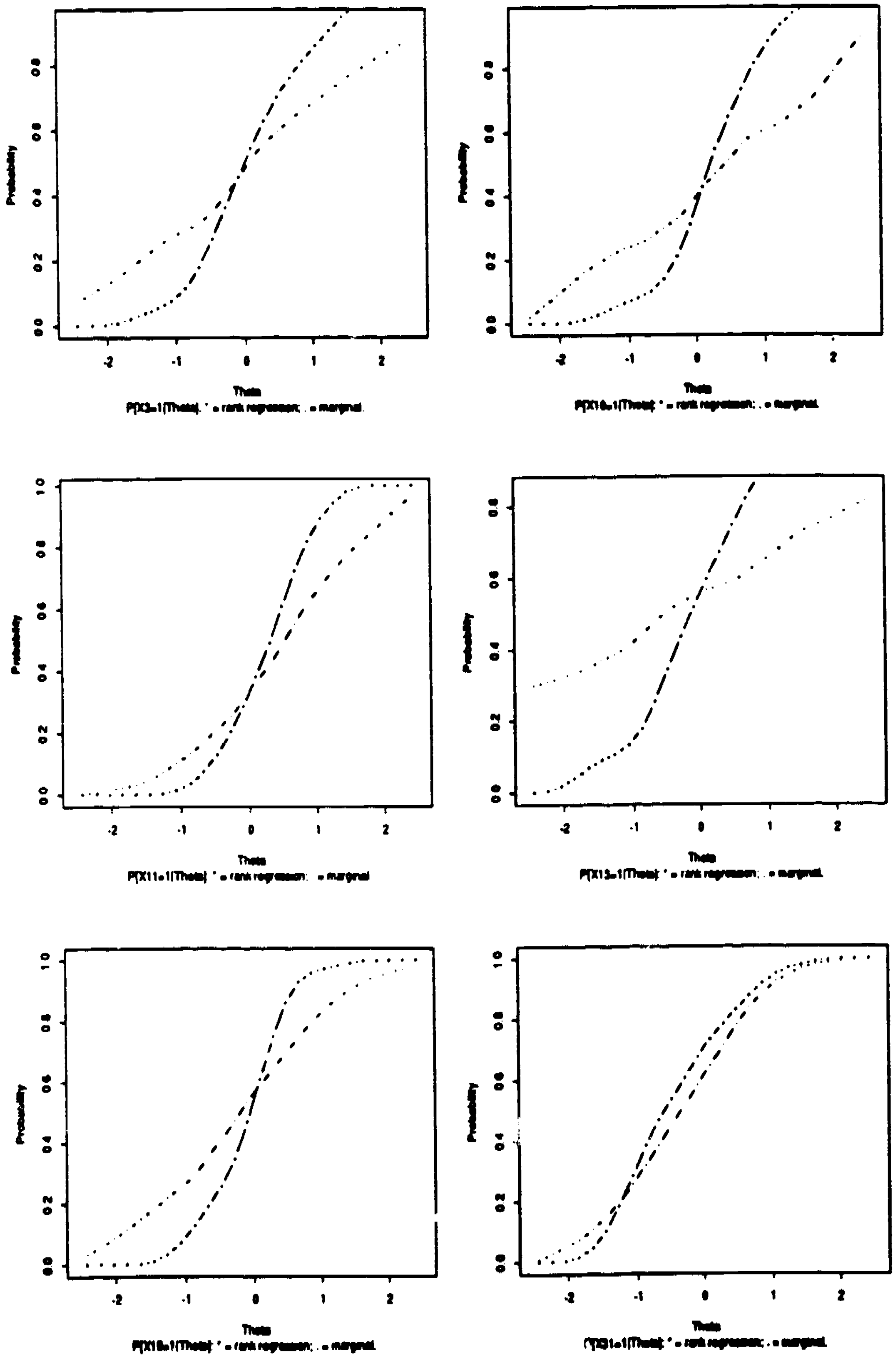
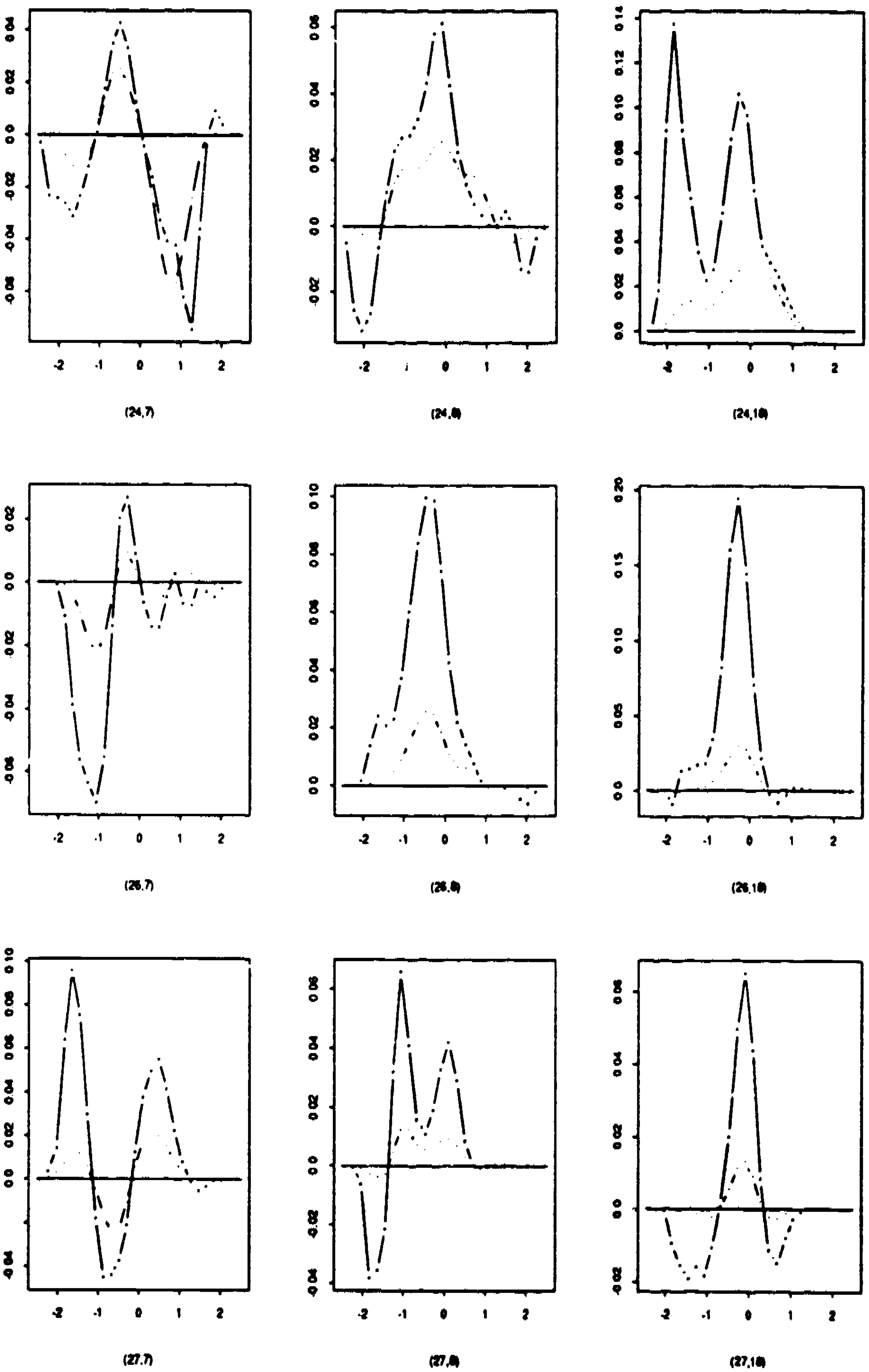


Figure 2: One-dimensional ICC's for two-dimensional data.



(24,7) (24,8) (24,10)

(26,7) (26,8) (26,10)

(27,7) (27,8) (27,10)

Figure 3: Unidimensional local covariance estimates for one-dimensional data.

both Figures 3 and 4 is probably related to the tendency for Cov $(X_j, X_j|X_+)$ to be negative in typical IRT data (see Junker, 1991a; as well as the discussion of the ASVAB illustration below).

Estimates of $C_J(\theta)$ for the two data sets are compared in Figure 5. Note how much higher $\hat{C}_J(\theta)$ is for the two-dimensional data set than for the one-dimensional data set. In this case the $\hat{C}_J(\theta)$'s are easy to compare, since they are based on data generated from similar models (both use logistic ICC's, and the one-dimensional parameters are the averages of the corresponding two-dimensional parameters) which differ only in latent space dimensionality.

The extent to which the unidimensional information and asymptotic MLE standard errors are too optimisic for the two-dimensional data set is illustrated in Figure 6. The graph on the left in Figure 6 is again $\hat{C}_J(\theta)$ for this data set. In the center and rightmost graphs in Figure 6, the uncorrected SE and information functions are plotted with the symbol "." and the corrected SE^* and information functions are plotted with "*". The vertical scale for the center graph ranges from 2.0 to 10.0 and for the right graph from 0.0 to 1.2.

Let us turn to another illustration. We have simulated $N = 500$ examinee response strings to three parameter logistic items whose parameters were estimated from the Arithmetic Reasoning and Auto Shop sections of the Armed Services Vocational Aptitude Battery (these are the same item parameters as used for the ASVAB simulations in Table 1 above). Figures 7 and 8 illustrate the uncorrected and corrected MLE standard errors for these ASVAB-AR and ASVAB-AS data sets. Once again the leftmost graph is our estimate of $C_J(\theta)$ and the middle and rightmost graphs contrast the (estimated) uncorrected SE and informations function (symbol ".") with the (estimated) $C_J(\theta)$-corrected quantities (symbol "*"). In both figures, most of the "action" in $\hat{C}_J(\theta)$ is in the range -0.1 to 0.3. The MLE standard errors hover around 2.0, which seems a bit high, but it is worth noting that the corrected standard errors are not much different from the uncorrected ones. The story is similar for the corrected and uncorrected information functions, which effectively range from about 0.0 to 1.0 or so. Thus, as one would hope for unidimensional data, $\hat{C}_J(\theta)$ did not "overcorrect" the unidimensional SE and information estimates.

The fact that $C_J(\theta)$ is negative for moderately low values of θ in Figures 7 and 8 is interesting: as observed above in Section 6, Cov $(X_i, X_j|X_+)$ tends to be negative for unidimensional data; see Junker (1991a). The presence of the nonzero guessing parameter tends to make low-ability responses independent (without having to condition on θ) and this mak negative values for Cov $(X_i, X_j|X_+)$ even more likely. (On the other hand, the extreme positive va s of $C_J(\theta)$ near $\theta = -3$ are probably due to poor estimates of $\lambda_j'(\theta)$.) The standard error and information graphs in Figures 7 and 8 suggest the uncorrected quantities are adequate for measuring variability of MLE ability estimates for these items.

Our last illustration is a simulated paragraph-comprehension data set. The test consists of eight 5-item testlets (this nice term comes from Wainer and Lewis, 1990). The item response functions were compensatory logistic, with the first five items loading only on θ_1 and θ_2, the next five items loading only on θ_1 and θ_3, the next five on θ_1 and θ_4, and so on, such that nine latent traits are needed to achieve local independence in this model. The discriminations a_j in each dimension were sampled from $N(1.20, (0.8)^2)$ until $0.5 \leq a_j \leq 3$ and difficulties b_j in each dimension were sampled from $N(-0.12, (0.84)^2)$ until $-4 \leq b_j \leq 4$. There were no guessing parameters. Recall from Example 5.1 that a test constructed in this way will be *essentially unidimensional*, $d_E = 1$,

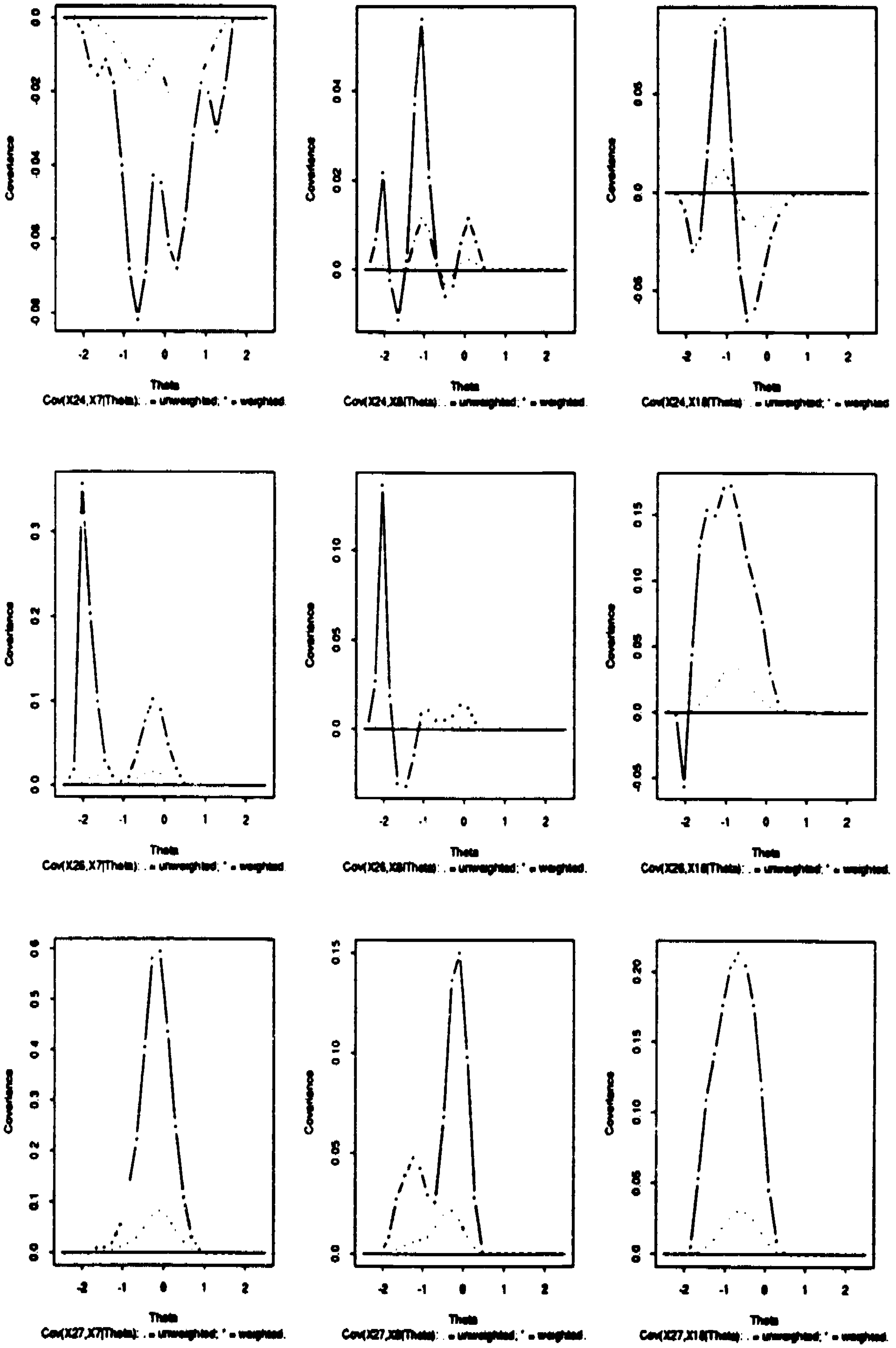


Figure 4: Unidimensional local covariance estimates for two-dimensional data.

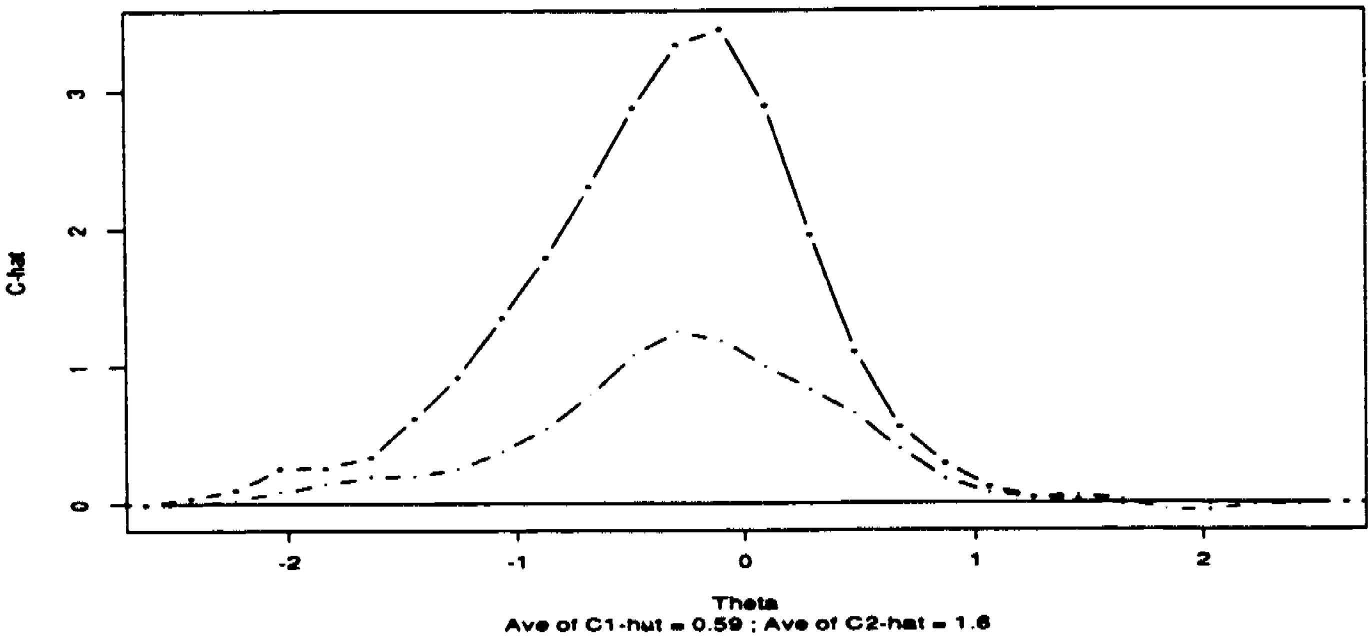


Figure 5: Comparison of $C_J(\theta)$ for one- and two-dimensional data.

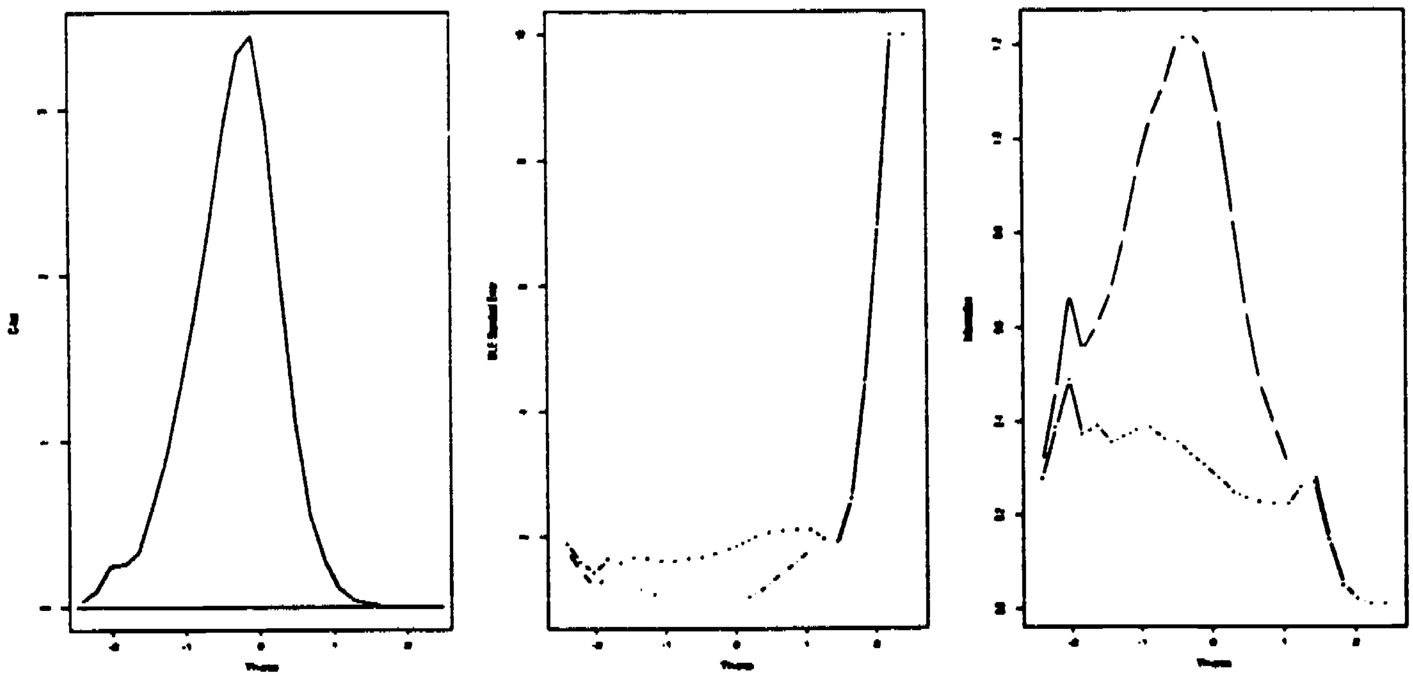


Figure 6: Accuracy of unidimensional ability estimates for two-dimensional data.

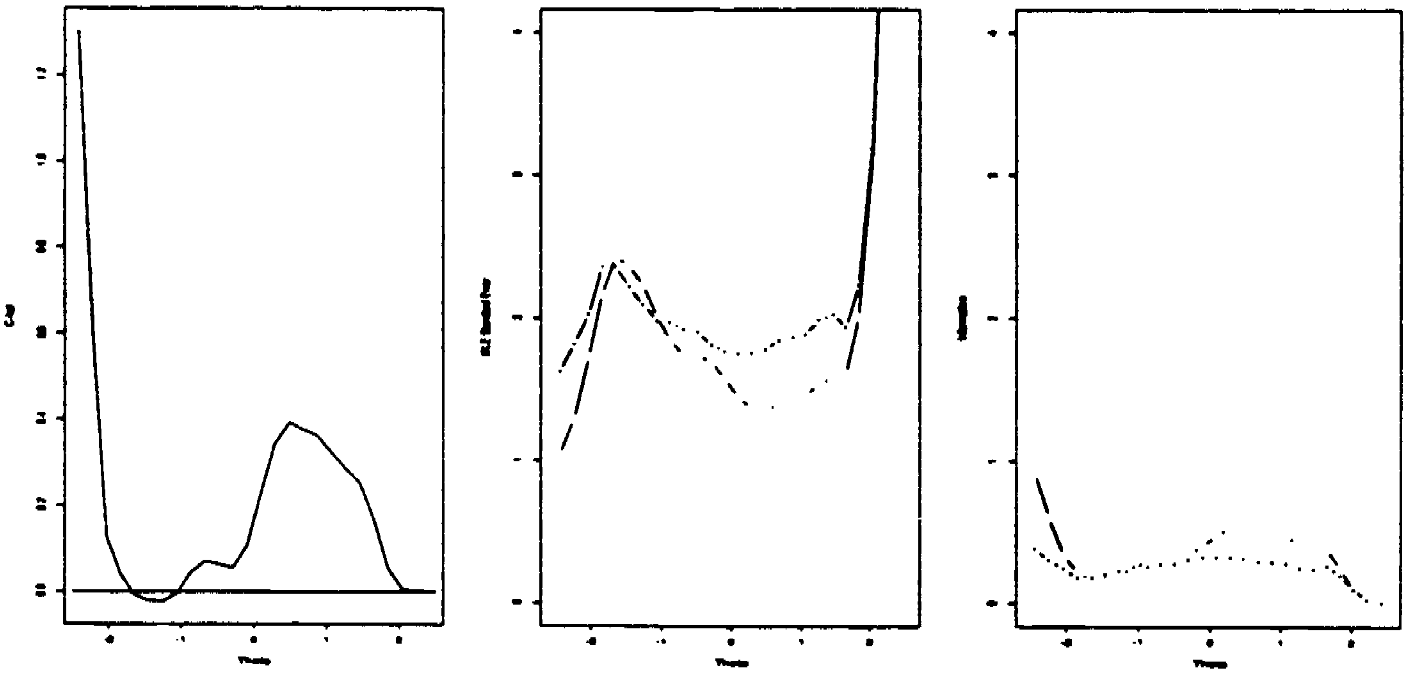


Figure 7: Accuracy of unidimensional ability estimates for simulated ASVAB-AR.

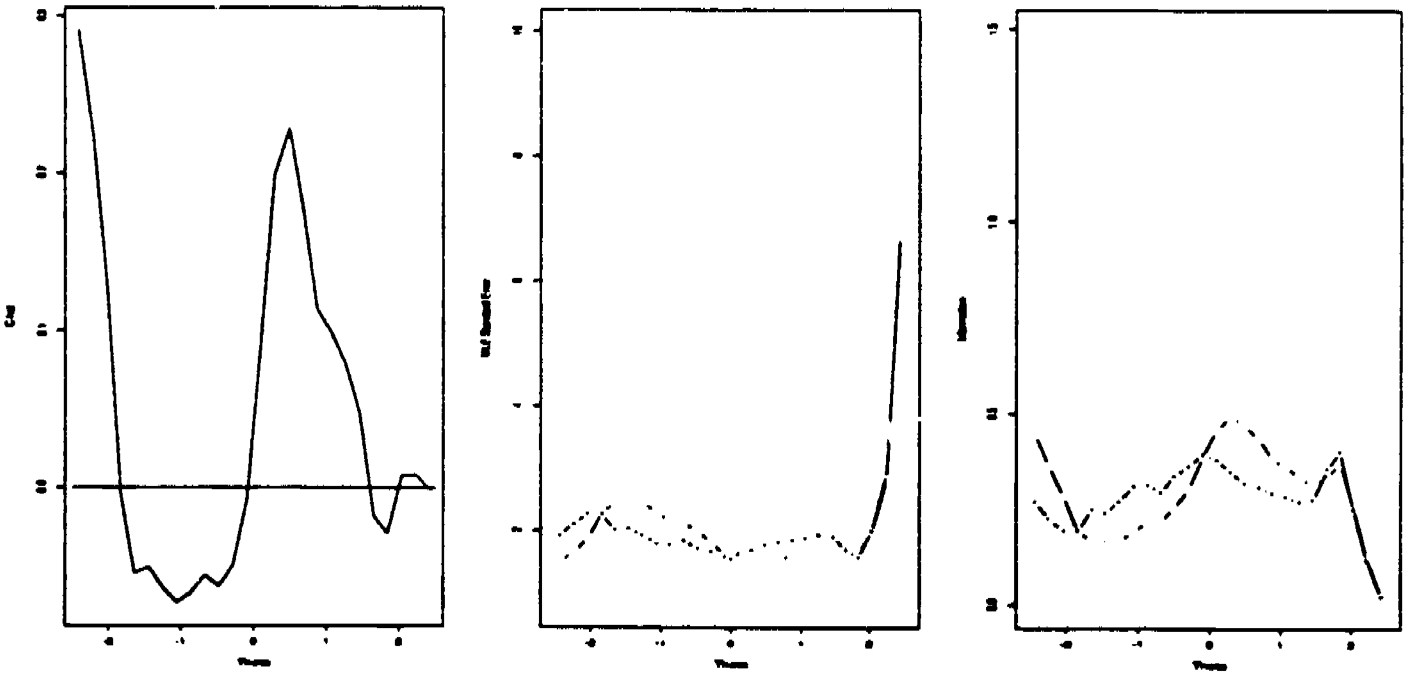


Figure 8: Accuracy of unidimensional ability estimates for simulated ASVAB-AS.

with respect to the dominant dimension θ_1. As usual, abilities in each dimension were generated to be i.i.d., $N(0,1)$.

In Figure 9, rank regression ICC estimates are compared, for a handful of items, with marginal θ_1-ICC's (here θ_1 is undoubtedly the ability "intended to be measured," though again a comparison with Wang's reference composite may be more appropriate). The rank regression ICC's match the marginal ICC's with respect to the dominant dimension θ_1 quite well. This suggests that for at least some $d_E = 1$ data sets, ICC's with respect to the dominant dimension *can* be recovered.

Figure 10 gives a graph o' $\hat{C}_J(\theta)$ based on these ICC estimates, and compares uncorrected (symbol ".") and corrected (symbol "*") measures of the variability of MLE ability estimates based on the rank regression ICC's. Here $\hat{C}_J(\theta)$ hovers between about 0.10 and 0.25, the standard errors hover just below 2.0, and the information functions live mostly between 0.2 and 0.4. Though there clearly would be some gain in employing a multidimensional model for this type of data, it is debatable whether it would be worthwhile, especially if the desire is to measure the dominant dimension only.

Comparing especially Figures 6, 7, 8 and 10, it appears that $\hat{C}_J(\theta)$ is a promising local index of unidimensionality. Clearly $\hat{C}_J(\theta)$ depends heavily on the local behavior of ICC's with respect to the dominant dimension being measured by the test, and especially on item parameters such as discrimination and guessing. Much more work needs to be done to understand this sensitivity and distinguish it from sensitivity to true multidimensionality in the data. It would also be interesting to run parallel studies of $\hat{C}_J(\theta)$ and the $\hat{\epsilon}_J$ index of Section 6, to see if they detect the same, or different, features of multidimensionality in item response data. Ultimately our goal is a *prescriptive* one: *do* use MLE, *don't* use MLE, *do* trust asymptotic normality, etc., depending on the size(s) of the indices. The work reported here suggests that such a goal should eventually be achievable.

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, **62**, 547–554.

Albert, J. H. (1991). Bayesian estimation of normal ogive item response curves using Gibbs sampling. Paper presented at the Workshop on Bayesian Computation via Stochastic Simulation, Ohio State University, Columbus Ohio, February 15–17.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in Lord, F. M. and Novick, M. R. (1968). *Stat'stical Theory of Mental Test Scores.* Addison-Wesley. Reading, Massachusetts.

Chung, H.-H. (1990). Asymptotic posterior normality of IRT models. Paper presented at the ONR Contractors' Meeting on Model-Based Psychological Measurement, Portland State University, Portland Oregon, June 16–19.

Clarke, B. S. and Junker, B. W. (1991). Inference from the product of marginals of a dependent likelihood. Submitted to the *Journal of the American Statistical Association.*

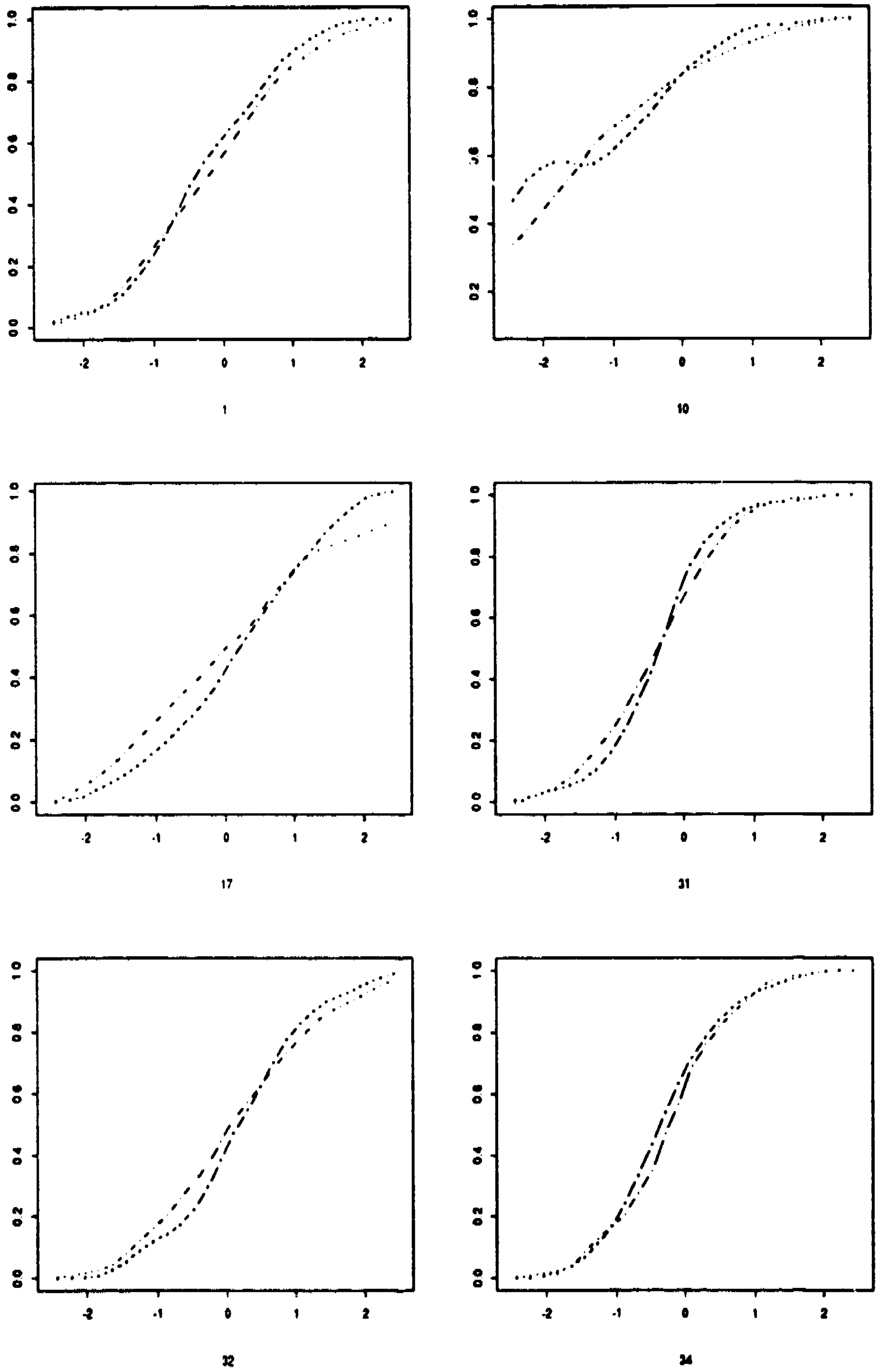


Figure 9: One-dimensional ICC's for simulated paragraph comprehension test.

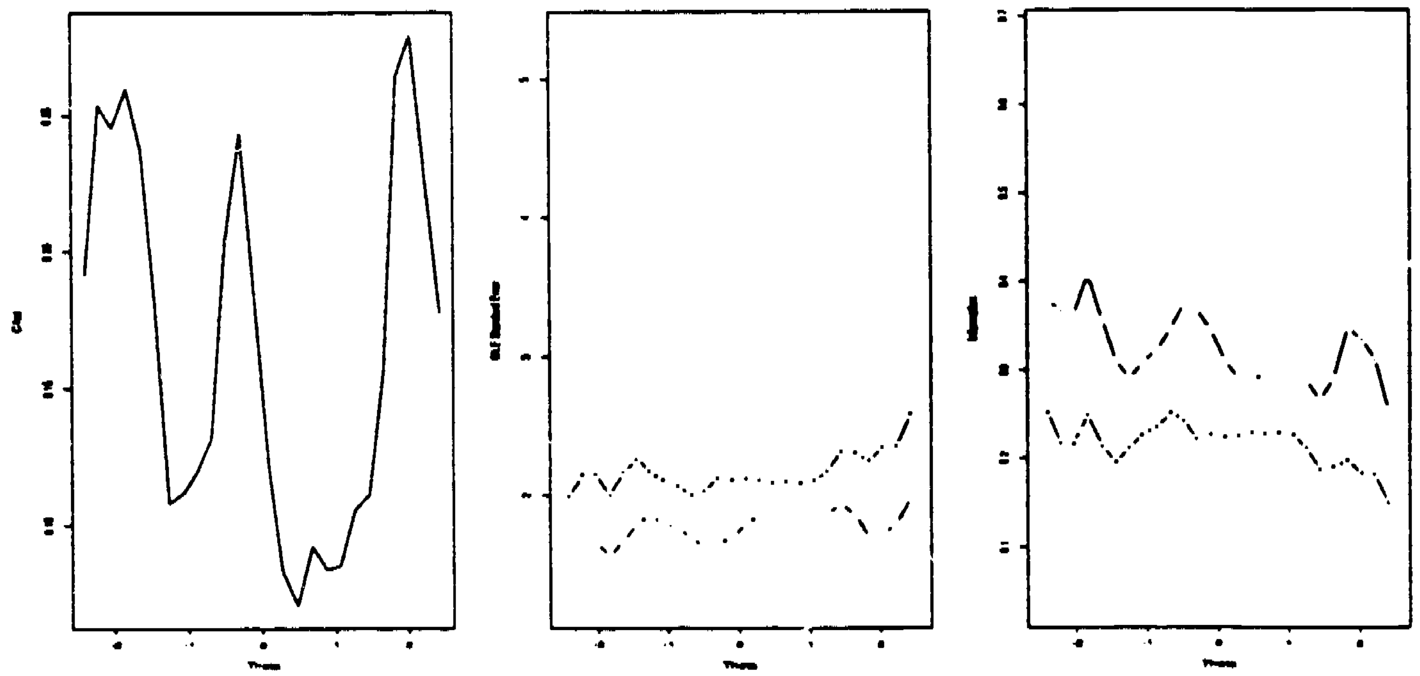


Figure 10: Accuracy of unidimensional ability estimates for simulated paragraph comprehension test.

Cox, J. T. and Grimmett, G. (1984). Central limit theorem for associated random variables and the percolation model. *Annals of Probability*, **12**, 514–528.

Csiszar, I. (1975). Information type measures of difference of probability distributions and direct observations. *Studia Scientiarum Mathematicum Hungarico*, **2**, 299–318.

Drasgow, F. and C. K. Parsons (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189–199.

Dvoretzky, A. (1972). Asymptotic normality for sums of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Probability and Statistics*, Volume II, 513–535.

Gibbons, R. D., Bock, R. D. and Hedeker, D. R. (1989). *Conditional dependence*. Final Research Report. Office of Naval Research and Illinois State Psychiatric Institute. Chicago, Illinois: University of Illinois at Chicago.

Holland, P. W. and Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent trait models. *Annals of Statistics*, 14, 1523–1543.

Iosifescu, M. and Theodorescu, M. (1969). *Random processes and learning*. New York: Springer-Verlag.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika*, 51, 357–373.

Junker, B. W. (1988). *Statistical aspects of a new latent trait model.* Ph.D. dissertation, Department of Statistics, University of Illinois. Champaign, Illinois.

Junker, B. W. (1991a). Conditional association, essential independence and monotone unidimensional item response models. Submitted to the *Annals of Statistics.*

Junker, B. W. (1991b). Essential independence and likelihood-based ability estimation for polytomous items. To appear, *Psychometrika*, 56.

Nandakumar, R. (1987). *Refinement of Stout's procedure for assessing latent trait unidimensionality.* Ph.D. dissertation, School of Education, University of Illinois. Champaign, Illinois.

Nandakumar, R. (1989). An improved statistical test for assessing essential unidimensionality in binary latent trait models. Submitted, *Journal of Educational Statistics.*

Nandakumar, R. (1991a). Assessing dimensionality of a set of items—comparison of different approaches. Paper presented at the Annual Meeting of the American Educational Research Association, April 1991, Chicago IL.

Nandakumar, R. (1991b). Traditional dimensionality vs. essential dimensionality. To appear, *Journal of Educational Measurement*, 28.

Newman, C. M. and Wright, A. L. (1982). Associated random variables and martingale inequalities. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **59**, 361–371.

Ramsay, J. O. (1990). A kernel smoothing approach to IRT modeling. Talk presented at the Annual Meeting of the Psychometric Society at Princeton New Jersey, June 28–July 1, 1990.

Reckase, M. (1990). Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests. Paper presented at the Annual Meeting of the American Educational Research Association, April 1990, Boston MA.

Spray, J. A. and Ackerman, T. A. (1987). The effect of item response dependency on trait or ability estimation. *ACT Research Report Series* #87-10. American College Testing Program. Iowa City, Iowa.

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

Stout, W. F. (1990a). Latent ability multidimensionality and an aymptotic item response theory modeling approach. Paper presented at the Annual Meeting of the American Educational Research Association, April 1990, Boston MA.

Stout, W. F. (1990b). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293–326.

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

Wainer, H. and Lewis, C. (1990). Towards a psychometrics for testlets. *Journal of Educational Measurement*, **27**, 1–14.

Wang, M.-M. (1986). *Fitting a unidimensional model to multidimensional item response data.* Paper presented at Office of Naval Research Model-Based Measurement Contractors' Meeting, Knoxville, TN, April 28, 1986.

Wang, M.-M. (1987). *Estimation of ability parameters from response data that are precalibrated with a unidimensional model.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC, April 22, 1987.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B*, 31, 80-88.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125–145.