ADAPTIVE TESTING, INFORMATION AND THE PARTIAL CREDIT MODEL: A SIMULATION STUDY

Raymond J. Adams

Centre for the Study of Higher Education
University of Melbourne
July 1987

CONTENTS

List of Tables

List of Figures

1. INTRODUCTION

In adaptive testing each individual is tested with a set of items that is selected to match his/her estimated ability[1] at the time of testing. In its most sophisticated form an adaptive test is interactively administered by a computer that scores the test and uses the individual's pattern of correct and incorrect responses to select new items from an item bank. In this procedure items are selected during the process of administering a test (rather than as a part of a predetermined sequence) so that the items administered to each individual are appropriate in difficulty for that individual. The result is a test that is matched to each individual's ability and the test is neither too easy nor too difficult. In this form adaptive testing requires application of both computer technology and latent trait theory.

The possibilities of using a latent trait model and computer technology to adapt a test for an individual on the basis of item difficulty is an area that has been covered extensively in psychometric research (for example, Weiss and Betz, 1972; Wood, 1972; Weiss, 1982). Over the past 10 to 15 years a variety of studies have demonstrated the possibility of achieving the same precision of measurement with fewer items (improved efficiency), and of obtaining smaller errors of measurement with the same number of items (improved precision) when items are selected on the basis of their appropriateness for individuals (see, Weiss, 1982; Weiss and Kingsbury, 1984). After a major three year

[1] Adaptive testing can be applied in any situation where the aim is to locate an individual on an underlying continuum or variable. However, for simplicity the word 'ability' will be used throughout this report.

project at the University of Minnesota, Weiss (1980) concluded that:

> In a variety of applications to the problem of achievement testing -- including measuring achievement with a large uni-dimensional item pool, measuring achievement levels in a number of specific content domains, and measuring achievement against a defined mastery criterion -- adaptive testing techniques using ICC theory can substantially reduce the number of items required in an achievement test without reducing the quality of measurements. Adaptive testing can improve the quality of achievement measurements in terms of both precision and validity while reducing the number of items required. (Weiss,1980:8)

To date however, these studies have been restricted to items based on dichotomous (correct/incorrect) scoring, a format that has become dominant through the widespread use of objectively scored paper and pencil tests. Although it has been mentioned in the literature (for example Bejar, 1976; McBride, 1979; Wood, 1973) the use of polychotomously scored items in adaptive testing does not appear to have been explored in any detail. A number of psychometric models are now available (Andrich, 1978; Bock, 1972; Masters, 1982; Samejima, 1969) that could be employed for adaptive testing with polychotomously scored items. The implementation of an adaptive procedure using one of these models is likely to further increase precision and efficiency by extracting information from partial knowledge exhibited in students' answers. It will also allow adaptive testing to be used with a greater range of item types; including likert style attitude items, item clusters, interactive items and items that use scoring that includes credit for partial understanding or partial completion.

In this report we aim to begin work on extending adaptive testing to include the Rating Scale Model (RSM) (Andrich, 1978) and Partial Credit Model (PCM) (Masters, 1982), two Rasch models that allow items to be scored in ordered categories. These models share statistical characteristics that enable the separation of person and item parameters and consequently they have sufficient statistics that are based on simple counts of objectively defined events (Masters and Wright, 1984). Parameter estimation with these models is relatively simple and could feasibly be applied in an interactive testing session with a micro-computer.

ELEMENTS OF ADAPTIVE TESTING

Item Banking

There are four key elements in any adaptive testing procedure. The first is a bank of test items with known characteristics, from which individually designed tests can be constructed. A latent trait model must be used to develop and maintain this bank of test items. Masters (1984) and Masters and Evans (1986) have shown how the partial credit model can be used to bank various types of polychotomously scored items. The rating scale model, which is perhaps best suited to attitude measurement, has not yet been used for item bank development but as a special case of the partial credit model it would seem almost obvious that item banking is possible with this model.

Test Scoring

Secondly, tests that differ from individual to individual must yield a score on a common scale. After calibrating an item bank the item parameters are treated as known and for each individual who takes the test an ability estimate can be calculated based on his/her responses to any subset of the items in the bank. This ability estimate can be used as a score because it reflects the location of the individual on the underlying continuum and is directly comparable to any other ability estimates based on tests built from the same item bank.

In adaptive testing, ability estimation (scoring) is usually undertaken with maximum likelihood or Bayesian approaches (Owen, 1969, 1975). In Bayesian estimation a prior ability distribution for each individual is hypothesized. After the administration of an item with known characteristics, the individual's response is used to modify that distribution to determine a posterior distribution. The mean of the posterior distribution becomes the updated ability estimate (score) and the standard deviation is the standard error of that estimate. In the Bayesian procedure the individual's ability is estimated sequentially, beginning with a prior estimate and updating that estimate item by item. The test score is the final updated ability estimate after the last item has been administered. Because of its sequential nature this method is order dependent (i.e., two individuals can be administered the same items and make the same responses but if the item order is different then their test scores may differ).

An alternative to Bayesian estimation is maximum likelihood estimation. In adaptive testing, where the item parameters are assumed to be known, it is relatively easy to use a Newton-Raphson maximum likelihood procedure to estimate the individual's ability on the basis of their previous responses. Maximum likelihood procedures make no assumptions regarding the individual's prior ability or the distribution of the attribute.

A number of studies have been undertaken to examine the relative merits of Bayesian and maximum likelihood based adaptive testing (e.g., McBride, 1976; Rosso & Reckase, 1981; Weiss, 1980; Weiss & McBride, 1983). The results of these studies have been mixed, but in general it would appear that maximum likelihood estimation is more accurate under a variety of conditions. While the Bayesian estimation can usefully employ prior knowledge of the individual's ability in the prior density function, there is a problem when the prior knowledge is not available or is inaccurate.

> This result points out the importance of the prior to the Bayesian procedure. An innacurate prior can effect the ability estimates. Since knowledge of the prior is often not available this procedure could result in biased estimates of ability. It thus seems that the maximum likelihood procedure is the procedure of choice if an adequate prior is not available. (Rosso and Reckase,1981: 10)

This result was supported by Weiss and McBride (1983) who found that Owen's Bayesian adaptive testing method did not provide measurements that were unbiased and equally precise at different levels of the ability continuum, except under the unrealistic condition of an extremely accurate prior. In this study we will assume no prior knowledge of individuals' ability and we will use maximum likelihood estimation.

Item Selection

A third key element in adaptive testing is the method used to select items from the bank for administration -- the item selection algorithm. The most common item selection algorithm applied with maximum likelihood estimation relies on the use of item information functions (Birnbaum, 1968) which indicate the amount of information that an item can convey about a particular ability level. Along with test information, which is the sum of the item information functions for the items in the test, item information has proved an important part of test development with latent trait models and with adaptive testing in particular.

Item selection is usually performed in the following way: Based on an individuals' responses to previous items, a maximum likelihood ability estimate is made. Given this ability estimate, the item information provided by each remaining item is calculated. The item providing the most information is then selected for administration. Selecting the item that provides the most information at an estimated ability is equivalent to trying to minimize the standard error of the post item, ability estimate.

The information function is well understood for the commonly used two and three parameter logistic (see, Birnbaum, 1968; Lord, 1980) and the dichotomous Rasch model. In the Rasch model the simplicity of the item information function has meant that its use has been largely implicit. For more complex latent trait models, that allow scoring in a number of ordered categories, the information functions are not well understood. Such an examination is necessary if adaptive testing algorithms are to be developed for these models. In chapter 2 RSM and the PCM are introduced and their information functions are discussed along with implications for adaptive testing.

Test Termination

The final element of adaptive testing is the rule used to determine when testing should be stopped -- the test termination criteria. Computerized adaptive testing is generally terminated at a fixed test length, which is likely to be somewhat shorter than for a conventional test, or after measurement error has decreased to some prespecified level which ensures equi-precise measurement over the ability range of interest. i.e., individuals of differing levels of ability have their ability estimated with equal accuracy. The consequences of a variety of stopping criteria are examined in Chapter 3 where a set of simulations using the RSM and the PCM are reported.

FOCUS OF THIS STUDY

In this report we focus on two major issues. The first is the nature of the information functions for the RSM and the PCM and the likely implications for adaptive testing. In chapter 2 we show that the information functions for these models behave in a manner that makes the implications for item bank construction and adaptive testing less than obvious.

In chapter 3 we report on some simulations with RSM and PCM item banks each with items scored in three response categories. These simulations provide some basic data on adaptive testing with these models.

2. THE MEASUREMENT MODELS AND THEIR INFORMATION FUNCTIONS

The models used in this study are members of the Rasch family of measurement models. As such they share "...a fundamental statistical characteristic -- separable person and item parameters and hence sufficient statistics." (Masters and Wright, 1984: 529). The fundamental building block of all of these models is Rasch's simple logistic model (Rasch, 1960) which can be used when items are scored dichotomously. Under this model the probability of an individual with ability β scoring x_i on item i with the single item parameter δ_{i1} is given by:

$$P_{x_i}(\beta) = \frac{\exp x_i(\beta - \delta_{i1})}{1 + \exp(\beta - \delta_{i1})} \qquad x_i = 0,1 \qquad (2.1)$$

The partial credit model (Masters, 1982) is an extension of the simple logisitic model that can be used with response formats that employ more than two ordered performance or response categories. Wright and Masters (1982) and Masters and Wright (1984) describe how the partial credit model can be built from the multiple application of the simple logistic model.

If item i has m_i+1 ordered response categories then the probability of an individual with ability β responding in category x_i to item i is given by:

$$P_{x_i}(\beta) = \frac{\exp \sum_{j-0}^{x_i} (\beta - \delta_{ij})}{\sum_{k-0}^{m_i} \exp \sum_{j-0}^{k} (\beta - \delta_{ij})} \qquad x_i=0,1,\ldots,m_i \qquad (2.2)$$

where item i is described by parameters $\delta_{i1},\ldots\ldots,\delta_{im}$ and

$$\exp \sum_{j-0}^{0} (\beta - \delta_{ij}) \equiv 1$$

The rating scale model (Andrich, 1978) is a special case of the partial credit model that places constraints on the possible values of the item parameters $\delta_{ij,}$ so that they can be reparameterized as δ_{ij}=$(\delta_i+\tau_j)$. In this case the number of categories is equal for all items and we can write m_i=m for all i. When this model is applied to the analysis of a rating scale, a location parameter, δ_i, is estimated for each item i, and m response 'thresholds' τ_1, $\tau_2,\ldots,\tau_m$, are estimated for the m+1 response alternatives. Model (2.2) can then be expressed as:

$$P_{x_i}(\beta) = \frac{\exp \sum_{j-0}^{x_i} (\beta - (\delta_i+\tau_j))}{\sum_{k-0}^{m} \exp \sum_{j-0}^{k} (\beta - (\delta_i+\tau_j))} \qquad x_i=0,1,\ldots,m \qquad (2.3)$$

where $\exp \sum_{j-0}^{0} (\beta - (\delta_i+\tau_j)) \equiv 1$ and $\sum_{j-1}^{m} \tau_j = 0$

INFORMATION FUNCTIONS

In adaptive testing one of the most important and useful latent trait concepts is test and item information. Given a set of test items and a vector $\mathbf{V}=(x_1,x_2,\ldots\ldots,x_L)$, of responses to those items, Birnbaum (1968: 453) defined the test information of any given scoring formula, $x=x(\mathbf{V})$, based on any given scoring model as:

$$I(\beta,x) = \frac{1}{\mathbf{var}(x|\beta)} \cdot \frac{[\partial \mathbf{E}(x|\beta)]^2}{\partial\beta} \qquad (2.4)$$

This definition of information was chosen by Birnbaum because it expressed the precision of interval estimation based on the given test and scoring formula. Lord (1980) noted that the numerator of this function is the squared slope of the regression of the test score on β, and the denominator is the conditional variance of the test score. Therefore, information increases as the slope of the expected score increases (i.e., as the score becomes more sensitive to changes in β), and as the variance in the score becomes smaller. Information when defined in this way has proven to be a useful measure of the accuracy of a test at various levels of β.

For the partial credit model defined in (2.2) the expected test score (regression of the test score on β) is given by:

$$\mathbf{E}(x|\beta) = \sum_{i=1}^{L} \sum_{k_i=0}^{m_i} k_i P_{k_i}(\beta)$$

and therefore,

$$\frac{\partial}{\partial\beta}\mathbf{E}(x|\beta) = \sum_{i=1}^{L} \frac{\partial}{\partial\beta} \sum_{k_i=0}^{m_i} k_i P_{k_i}(\beta)$$

$$= \sum_{i=1}^{L} \left[\sum_{k_i=1}^{m_i} k_i^2 P_{k_i}(\beta) - \left\{ \sum_{k_i=1}^{m_i} k_i P_{k_i}(\beta) \right\}^2 \right]$$

then by local independence,

$$\frac{\partial}{\partial\beta}\mathbf{E}(x|\beta) = \mathbf{var}(x|\beta) \qquad (2.5)$$

Substituting (2.5) into (2.4) gives I(β,x) = **var**(x|β), which corresponds to the reciprocal of the asymptotic variance for the unconditional maximum likelihood estimator of β (Wright and Masters, 1982: 82). This is

1982: 82). This is the usual concept of information developed by Fisher. In general $I(\beta, x)$ will be simply written as $I(\beta)$ with $x=x(\mathbf{V})$ implied.

Since the test information consists of additive components summed over the number of items it is possible to define <u>item information</u> as:

$$I_i(\beta) = \sum_{k=1}^{m_i} k_i^2 P_{k_i}(\beta) - \left(\sum_{k=1}^{m_i} k_i P_{k_i}(\beta)\right)^2 \qquad (2.6)$$

so that:

$$I(\beta) = \sum_{i=1}^{L} I_i(\beta)$$

The work of Birnbaum was based on items that could be scored 0 or 1, (Birnbaum, 1968: 397). When Samejima (1969) developed the graded response model she extended the definition of information to cover items scored in a number of graded categories.

Samejima defined the <u>item response information function</u> as:

$$I_{x_i}(\beta) = -\frac{\partial^2}{\partial\beta^2} \log P_{x_i}(\beta)$$

this definition can be expressed as:

$$I_{x_i}(\beta) = -\frac{\partial}{\partial\beta} \frac{P'_{x_i}(\beta)}{P_{x_i}(\beta)}$$

$$= -\frac{P''_{x_i}(\beta)P_{x_i}(\beta) - [P'_{x_i}(\beta)]^2}{[P_{x_i}(\beta)]^2}$$

$$= \left[\frac{P'_{x_i}(\beta)}{P_{x_i}(\beta)}\right]^2 - \frac{P''_{x_i}(\beta)}{P_{x_i}(\beta)} \qquad (2.7)$$

She then defined the item information function as the expected value of the item response information.

$$I_i(\beta) \;=\; \mathbf{E}\,[I_{x_i}(\beta)]$$

that is:

$$I_i(\beta) \;=\; \sum_{k=0}^{m_i} I_{x_i}(\beta)\, P_{x_i}(\beta)$$

and substituting from (2.7) gives:

$$I_i(\beta) \;=\; \sum_{k=0}^{m_i} \frac{P'_{x_i}(\beta)^2}{P_{x_i}(\beta)} \;-\; \sum_{k=0}^{m_i} P''_{x_i}(\beta)$$

To make some of the algebra easier the function:
$\Psi_\eta = \Psi_\eta(\beta) = \exp(\eta\beta - \delta_{i1} - \dots - \delta_{i\eta})$ with $\Psi_0 = 1$, can be defined.

Note then that $$\frac{\partial}{\partial\beta}\,\Psi_\eta(\beta) = \eta\Psi_\eta(\beta)$$

and that the PCM as defined in (2.2) can be expressed as:

$$P_{x_i}(\beta) \;=\; \frac{\Psi_{x_i}(\beta)}{\sum\limits_{k=0}^{m_i}\Psi_k(\beta)}$$

then

$$\frac{\partial}{\partial\beta} P_{x_i}(\beta) \;=\; \frac{x_i\Psi_{x_i}(\beta)\sum\limits_{k=0}^{m_i}\Psi_k(\beta) - \Psi_{x_i}(\beta)\sum\limits_{k=0}^{m_i}k\Psi_k(\beta)}{\left[\sum\limits_{k=0}^{m_i}\Psi_k(\beta)\right]^2}$$

$$= P_{x_i}(\beta) \left\{ x_i - \sum_{k=0}^{m_i} kP_k(\beta) \right\}$$

$$= P_{x_i}(\beta) \left\{ x_i - \mathbf{E}(x|\beta) \right\}$$

and

$$\frac{\partial^2}{\partial\beta^2} P_{x_i}(\beta) = \frac{\partial}{\partial\beta} P_{x_i}(\beta) \left\{ x_i - \sum_{k=0}^{m_i} kP_k(\beta) \right\}$$

$$= P'_{x_i}(\beta) \left[x_i - \sum_{k=0}^{m_i} kP_k(\beta) \right] - P_{x_i}(\beta) \sum_{k=0}^{m_i} k\, P'_k(\beta)$$

$$= P_{x_i}(\beta) \left[x_i - \sum_{k=0}^{m_i} kP_k(\beta) \right]^2 - P_{x_i}(\beta) \sum_{k=0}^{m_i} \left\{ k \left[k - \sum_{k=0}^{m_i} kP_k(\beta) \right] P_k(\beta) \right\}$$

$$= P_{x_i}(\beta) \left[x_i - \sum_{k=0}^{m_i} kP_k(\beta) \right]^2 - P_{x_i}(\beta) \left[\sum_{k=0}^{m_i} k^2 P_k(\beta) - \left\{ \sum_{k=0}^{m_i} kP_k(\beta) \right\}^2 \right]$$

so $\sum_{x_i=0}^{m_i} P''_x(\beta) = 0$

then

$$I_i(\beta) = \sum_{x_i=0}^{m_i} \frac{[P'_{x_i}(\beta)]^2}{P_{x_i}(\beta)}$$

$$= \sum_{x_i-0}^{m_i} \frac{(P_{x_i}(\beta) [x_i - \mathbf{E}(x_i|\beta)] \quad)^2}{P_{x_i}(\beta)}$$

$$= \sum_{x_i-0}^{m_i} [x_i - \mathbf{E}(x_i|\beta)]^2 P_x(\beta)$$

$$- \mathbf{var}(x_i | \beta)$$

If $I(\beta)$, the test information is defined as the sum of the item informations then

$$I(\beta) = \sum_{i-1}^{L} I_i(\beta)$$

$$= \sum_{i-1}^{L} \mathbf{var}(x_i | \beta)$$

$$= \mathbf{var}(x | \beta) \quad \text{(by local independence)}$$

So both the Samejima and Birnbaum formulations coincide, and correspond to the reciprocal of the asymptotic variance of $\hat{\beta}$, the maximum likelihood estimator of β.

ITEM INFORMATION AND THE SIMPLE LOGISTIC MODEL

In examining the nature of the information function for these models it is convenient to begin with the simple logistic model (2.1). In this case:

$$I_i(\beta) = P_{x_i}(\beta) - P_{x_i}(\beta)^2$$
$$= P_{x_i}(\beta)(1 - P_{x_i}(\beta))$$

Since $0 \leq P_{x_i} \leq 1.0$, $I_i(\beta)$ attains a maximum of 0.25 at $\beta=\delta_{i1}$ and all information curves have the same shape. Figure 2.1 shows the

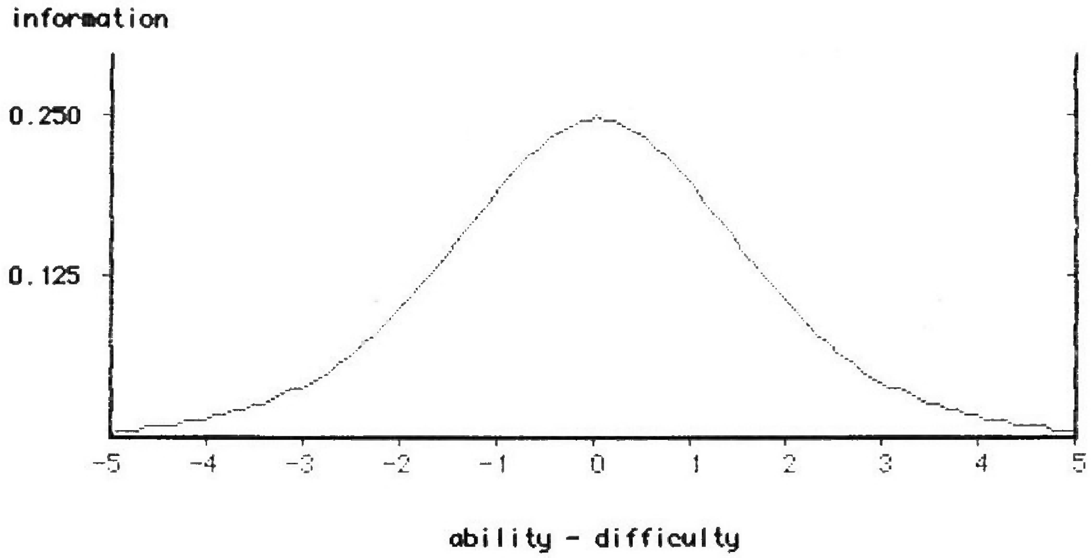

Figure 2.1 Information curve for a two category item

information curve as a function of β for δ_{11}-0. For items with different δ_{11} the curves will differ only in their location on the variable and that location can be unambiguously interpreted as the item difficulty. Consequently, when all items in an item bank are scored dichotomously the item with the difficulty paramter that most closely matches the current ability estimate maximizes the available information.

THREE RESPONSE CATEGORIES

For the PCM with m_i=2, (2.2) can be expressed as follows for x_i=0, 1 or 2 respectively:

$$P_0 = \frac{\Psi_0(\beta)}{\Psi_0(\beta)+\Psi_1(\beta)+\Psi_2(\beta)} \qquad (2.8)$$

$$P_1 = \frac{\Psi_1(\beta)}{\Psi_0(\beta)+\Psi_1(\beta)+\Psi_2(\beta)} \tag{2.9}$$

$$P_2 = \frac{\Psi_2(\beta)}{\Psi_0(\beta)+\Psi_1(\beta)+\Psi_2(\beta)} \tag{2.10}$$

and substituting (2.8), (2.9) and (2.10) into (2.6):

$$I_i(\beta) = \frac{\Psi_1 + 4\Psi_2}{\Psi_0+\Psi_1+\Psi_2} - \left[\frac{\Psi_1 + 2\Psi_2}{\Psi_0+\Psi_1+\Psi_2}\right]^2 \tag{2.11}$$

If we consider an item i, with fixed parameters then the value of β for which the information attains a maximum may be determined by differentiating (2.11) with respect to β.

$$\frac{\partial}{\partial\beta} I_i(\beta) = \frac{\Psi_1 - \Psi_1^2 + \Psi_1^2\Psi_2 - \Psi_1\Psi_2^2 + 8\Psi_2 - 8\Psi_2^2}{(\Psi_0+\Psi_1+\Psi_2)^3}$$

For $\beta=(\delta_{i1}+\delta_{i2})/2$, $\Psi_2 = 1$ therefore,

$$\frac{\partial}{\partial\beta} I_i(\beta) = \frac{\Psi_1 - \Psi_1^2 + \Psi_1^2 - \Psi_1 + 8 - 8}{(\Psi_0+\Psi_1+\Psi_2)^3} = 0$$

Consequently $\beta=(\delta_{i1}+\delta_{i2})/2 = \delta_{i\bullet}$ is a turning point. If $\beta=\delta_{i\bullet}$, is substituted into (2.11) we find that:

$$I_i(\delta_{i\bullet}) = \frac{\Psi_1 + 4}{\Psi_1 + 2} - \left[\frac{\Psi_1+2}{\Psi_1+2}\right]^2 = \frac{2}{2 + \exp((\delta_{i2}-\delta_{i1})/2)} \tag{2.12}$$

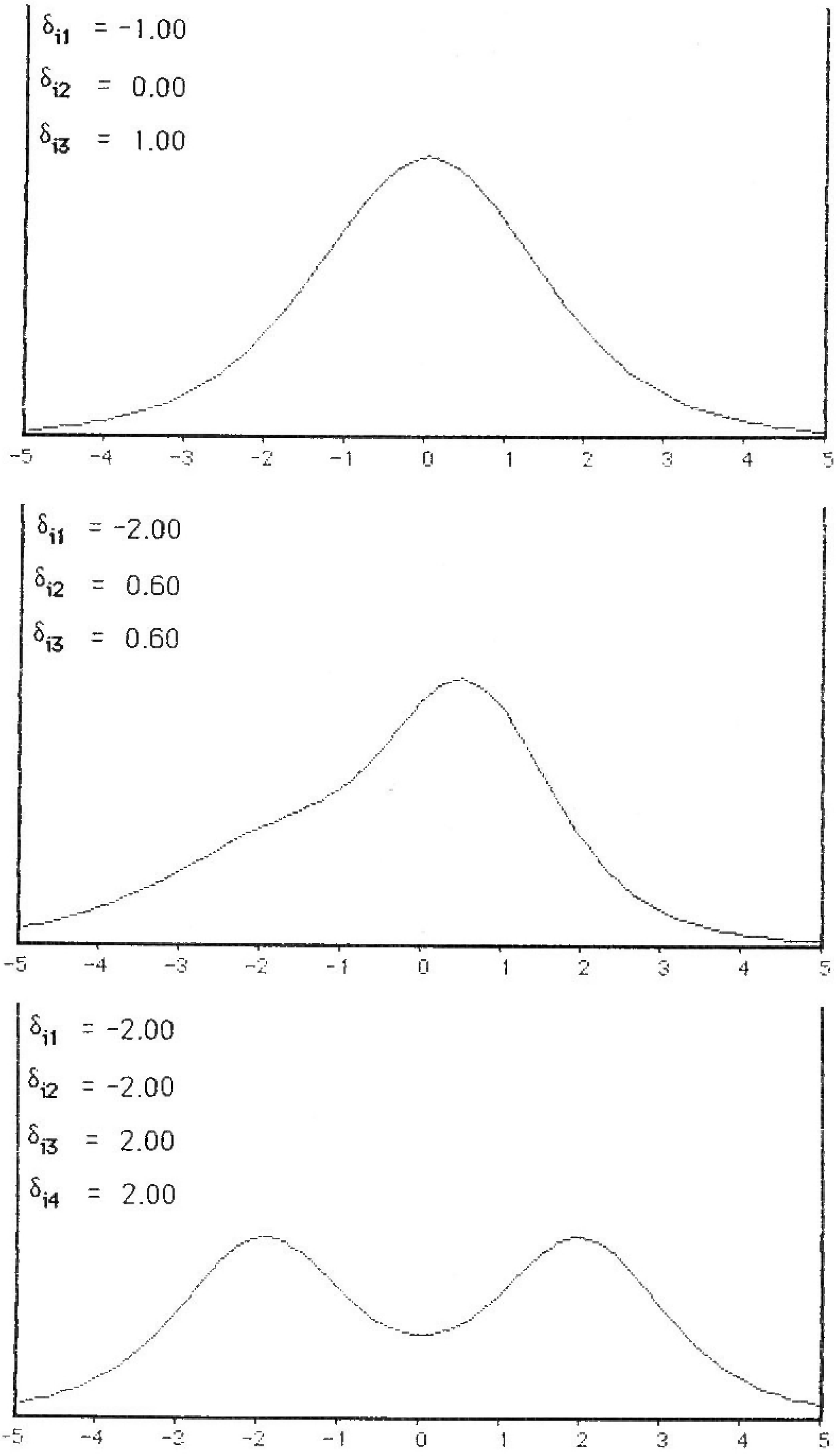

Figure 2.2 Some possible item information curves

So while $\beta=\delta_{i\bullet}$ will always be a turning point it may be a global or local, minimum or maximum. That is, depending on the values of δ_{i1} and δ_{i2} the shape of the information curve for a given item, i, depends on the difference between the item parameters δ_{i1} and δ_{i2} for that item. Three possibilities are:

- If $\delta_{i1}=\delta_{i2}$ then $I_i = 2/3$ and $\beta=\delta_{i\bullet}$ is the global maximum
- as $(\delta_{i2}-\delta_{i1}) \rightarrow \infty$, $I_i \rightarrow 0$ and $\beta=\delta_{i\bullet}$ is a minimum
- and as $(\delta_{i2}-\delta_{i1}) \rightarrow -\infty$, $I_i \rightarrow 1$ and $\beta=\delta_{i\bullet}$ is the global maximum

In summary, if δ_{i2} is much larger than δ_{i1} then the information function has a local minimum at $\beta=\delta_{i\bullet}$ and is bimodal with peaks at δ_{i1} and δ_{i2}. If δ_{i1} is greater than δ_{i2} then the information function has a maximum at $\beta=\delta_{i\bullet}$. As $\delta_{i2}-\delta_{i1}$ tends to negative infinity the function becomes more peaked. Unlike the dichotomous case of the model the information functions for items with different parameters no longer always have identical shapes since the shape of the curve is a function of $\delta_{i2}-\delta_{i1}$. For $\delta_{i2}-\delta_{i1}$ less than about 3.2 logits I_i is a maximum and for $\delta_{i2}-\delta_{i1}$ greater than 3.2 I_i is a local miminum.

THE GENERAL CASE (m_i>2)

For the general case of m_i+1 ordered categories in item i the algebra involved in an analytical examination of the information functions becomes messy and resorting to empirical examples is more useful. Figure 2.2 contains plots of the information curves for three items with parameters given by;

(i) $\delta_{i1}=-1$ $\delta_{i2}=0$ $\delta_{i3}=1$

(ii) $\delta_{i1}=-2$ $\delta_{i2}=0.6$ $\delta_{i3}=0.8$

(iii) $\delta_{i1}=-2$ $\delta_{i2}=-2$ $\delta_{i3}=2$ $\delta_{i4}=2$

For item (i) the information curve is symmetrical with a peak at $\beta=0$.

For item (ii) the curve is skewed slightly to the right and peaks at about $\beta=0.5$. For item (iii) the curve is bimodal with peaks at about $\beta=-2$ and $\beta=2$. Dodd and Koch (1985) noted that while the shape of the information curves can be quite different, the area under the curves (total information) is identical whenever the items have the same number of item parameters. Although not shown in their paper this can be easily demonstrated since:

$$\begin{aligned} \int_{-\infty}^{\infty} I_i(\beta)\, d\beta &= \Big[E(x_i \mid \beta) \Big]_{\infty}^{-\infty} \\ &= m_i \end{aligned}$$

It is interesting to note that this is an essential feature of Rasch measurement models. Although it is often stated that if items are to conform to the Rasch model they must have equal item discriminations, the varying shapes of the information curves shown above indicate that this will not generally be the case if item discrimination is defined as the slope of the expected item score (regression of item score on ability). To conform to the Rasch model all items with the same number of categories must contribute the same amount of total information and the slopes of the conditional probability curves, $P_x/(P_{x-1}+P_x)$, must be the same. This of course implies that the simple logistic model requires equal item discrimination since each item contains only one decision point.

The preceeding discussion shows that it is misleading to talk of a single value of item discrimination, unless a linear item characteristic curve is proposed, as in the classical test model. For most latent trait models the discriminating power of an item at a given ability level is a function of all item parameters. Examining the information curve is the natural way to determine these variations in discrimination power.

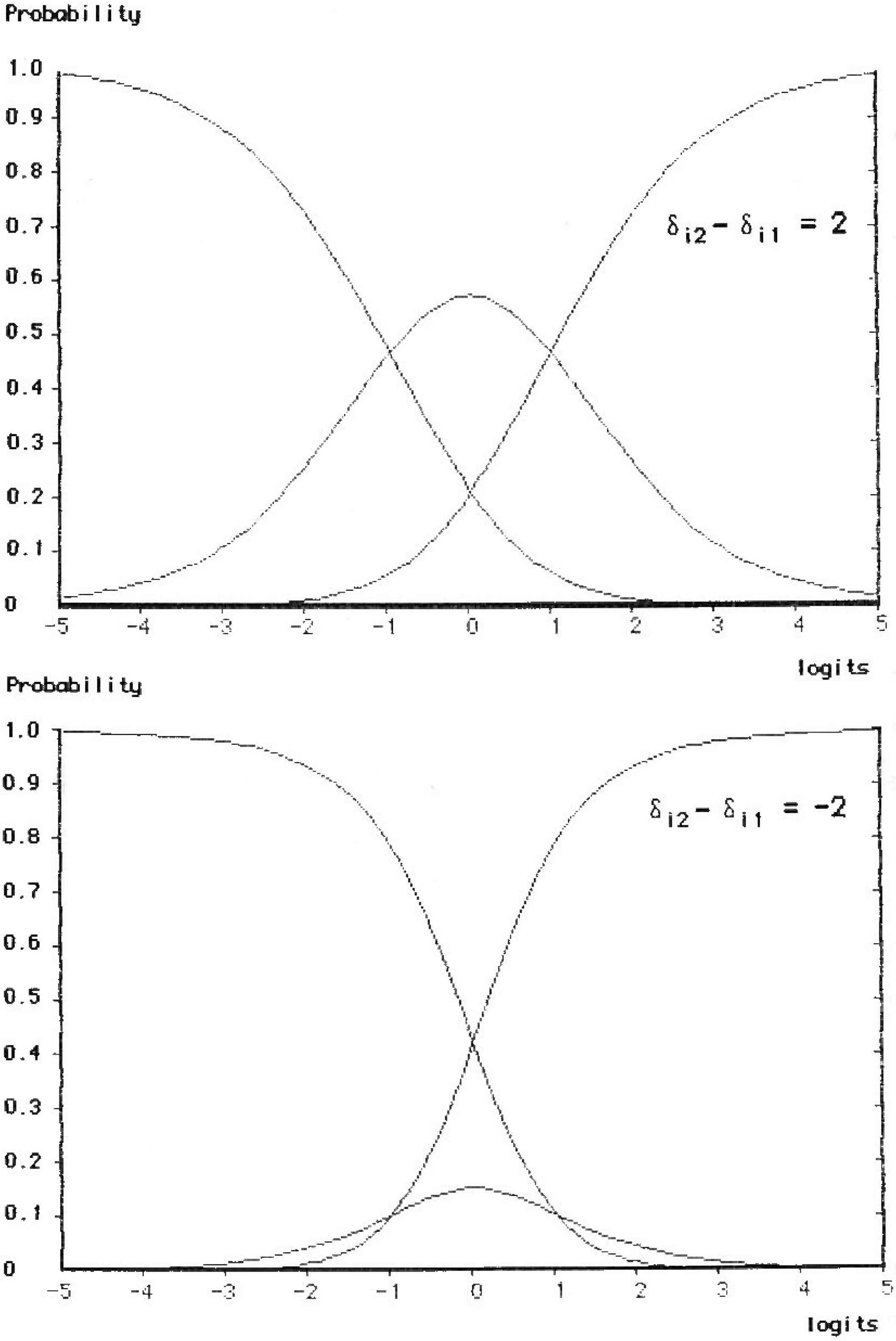

Figure 2.3 Item characteristic curves for two three category items

MAXIMIZING INFORMATION

The amount of information that an item provides at different levels of ability is a function of the item's sensitivity to differences in ability (discriminating power). This is illustrated by Figures 2.3, 2.4. and 2.5 These figures show the item characteristic curves, the expected score curves and the item information curves for two three category items with $\delta_{i2}-\delta_{i1}=2$ and $\delta_{i2}-\delta_{i1}=-2$ respectively. These figures illustrate how the height of the information curve is related to the slope of the expected score curve, which is the regression of the item score on β. Where the gradient of the expected score is steep a small change in ability results in a large change in the expected score of the student. When the expected score curve is less step the item is less sensitive to changes in ability.

The information curve (see Figure 2.5) for the item with $\delta_{i2}-\delta_{i1}=-2$ is considerably more peaked than the information curve for the item with $\delta_{i2}-\delta_{i1}=2$. If the corresponding item characteristic curves are examined it can be seen that the item with $\delta_{i2}-\delta_{i1}=-2$ is more sensitive than the item with $\delta_{i2}-\delta_{i1}=2$ to variations in β in the neighbourhood $\beta=0$. Figure 2.3 shows that most students with ability anywhere between $\beta=-0.5$ and $\beta=0.5$ are likely to score one for the item with $\delta_{i2}-\delta_{i1}=-2$, while for the item with $\delta_{i2}-\delta_{i1}=2$ students with β near 0.5 are likely to score 0 and students with β near 0.5 are likely to score 2.

While the item with $\delta_{i2}-\delta_{i1}=-2$ is more sensitive in the neighbourhood of $\beta=0$ the item with $\delta_{i2}-\delta_{i1}=2$ is more sensitive in the neighbourhoods of $\beta=-1.0$ and $\beta=1.0$. Again this can be seen by examining the item characteristic curves which indicate the sensitivity of the items to variations in β. For example, for β in the neighbourhood of -1.0 a

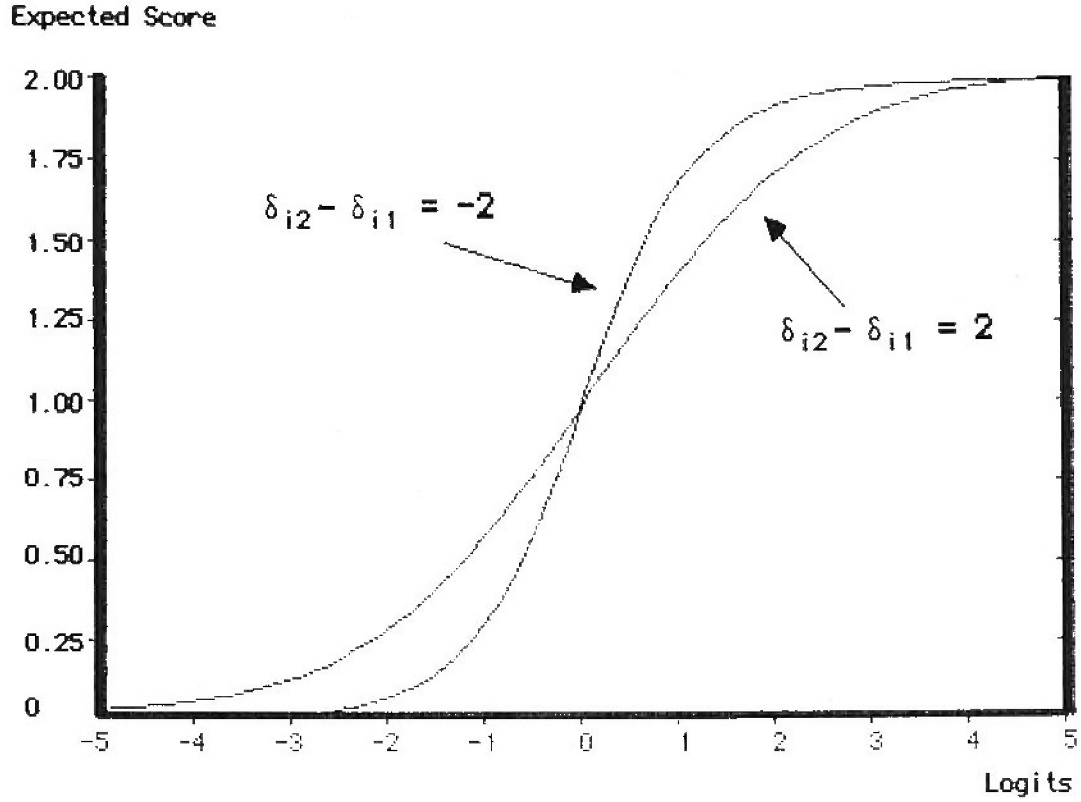

Figure 2.4 Expected score curves for two three category items

student will almost certainly score zero on the item with $\delta_{i2}-\delta_{i1}=2$ but for the item with $\delta_{i2}-\delta_{i1}=-2$ a student is most likely to score zero if $\beta<-1.0$ and one if $\beta>1.0$. Note that these two items provide the same total information and although the curves are different in shape they both have the same total area. This is reflected in Figure 2.4 as the item with $\delta_{i2}-\delta_{i1}=-2$ is more discriminating between $\beta=-1$ and $\beta=1$ and the other item is more discriminating elsewhere.

IMPLICATIONS FOR ADAPTIVE TESTING

When selecting items for administration in an adaptive test based on the dichotomous Rasch model ($m_i=1$) Figure 2.1 indicates that the item bank need only be searched for the item whose difficulty parameter most closely matches $\hat{\beta}$, the student's current ability estimate. In this case the selected item will be the item upon which the student has an expected

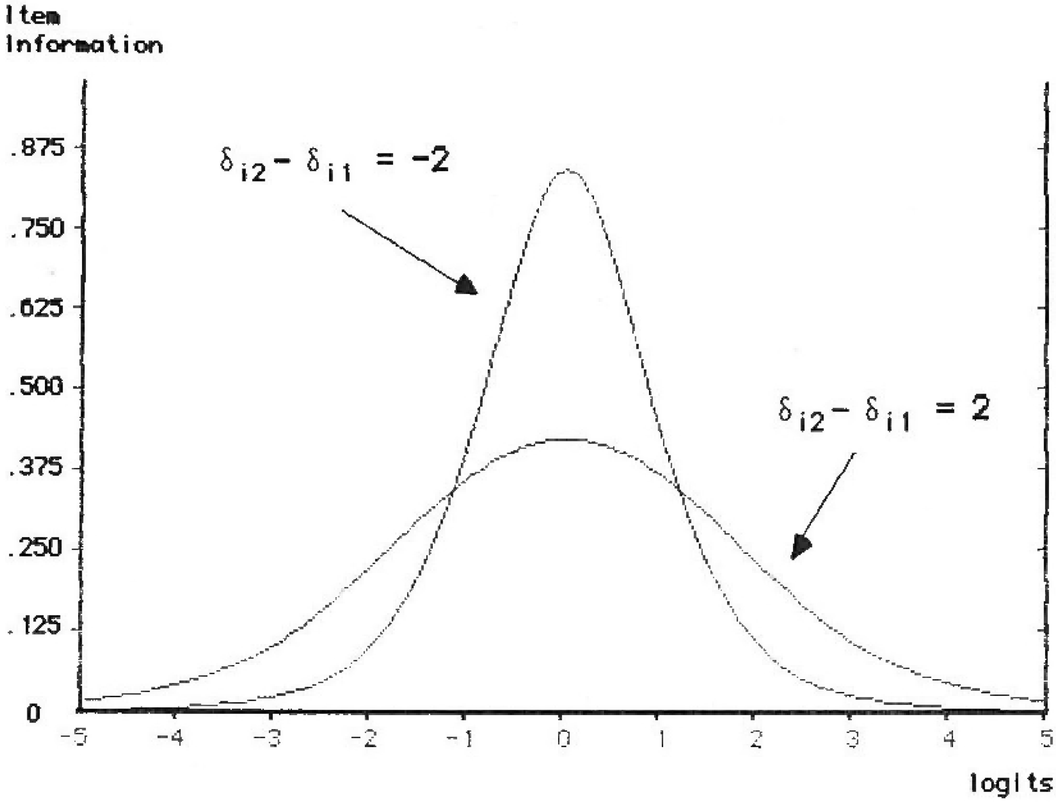

Figure 2.5 Information Curves for two three Category Items

score closest to 0.5. It would appear that there are three possible rules for item selection in adaptive testing when maximum likelihood estimation is used with the RSM and PCM:

(1) Select the item that most closely satisfies $E(x_i|\beta)=m_i/2$
(2) Select the item that most closely satisfies $\beta=\delta_{i\bullet}$
(3) Select the item for which $I_i(\beta)$ is a maximum

Previous examples and discussions indicate that these rules do not coincide in the general case of m_i+1 ordered categories. For the case $m_i=1$ all three conditions will be satisfied concurrently. For $m_i=2$ conditions (1) and (2) are satisfied at $\beta=\delta_{i\bullet}$ but (3) may not be satisfied. In fact for very large values of $\delta_{i2}-\delta_{i1}$ the information function may be a local minimum at $\beta=\delta_{i\bullet}$. For the case $m_i>2$ it is possible that none of the above conditions may be satisfied by the same item.

In the future it is expected that a study should be undertaken to determine which of these criteria (or perhaps some other) should be employed for item selection in an educational environment. On the basis of measurement precision criteria (3) would be the one of choice. However an examination of Figure 2.3 shows that items of high information may not be the most acceptable on educational grounds. For this item the score category one is not being used since the probability of scoring one is almost zero for all students. This means that students tend to get the item right or wrong and it may not be suited to partial credit scoring. Using three categories may have artificially inflated the discrimination power of the item at $\beta=\delta_{i\bullet}$.

DISCUSSION

When constructing an item the test developer makes a decision about the number of possible useful outcomes that the item can identify. In so doing the constructor decides upon the amount of information that the item can provide about the ability being measured. How these outcomes are defined then effects the location of the item parameters, δ_{ij} , and this determines the distribution of the available information over the variable of interest. This is in contrast to other popular latent trait models such as the two and three paramater logistic models, the graded response or nominal response models which allow the amount of information provided by an item to be determined empirically.

There does not appear to be any obvious rules to apply in determining an acceptable distribution of information for an item. If a developer defines three outcomes but one outcome is never very likely then the information function will be peaked and the item will be functioning as a dichotomy. The acceptability of this can only be determined by the developer in terms of whether the item is functioning as intended. It is possible that the substantive importance of an outcome will mean that it should be retained even when only few outcomes of the type occur.

In most instances it is expected that items regarded as acceptable by developers will provide maximum information at $\delta_{i\bullet}$. Consequently, in the following simulation we consider only items with an information maximum at $\delta_{i\bullet}$.

3. SIMULATIONS

In this chapter four sets of simulations are undertaken as a preliminary examination of adaptive testing with the rating scale and partial credit models. These simulations were undertaken to determine the recovery power of adaptive tests using various item bank sizes and test termination criteria.

The four sets of simulations were all undertaken with items scored in three categories only (i.e., $m_i=2$ for all i). In the first and second sets, $\delta_{i2}-\delta_{i1}$ was fixed at 1.0 for all items. This restricts the partial credit model to the rating scale model (Andrich, 1978). In the first set fixed length stopping rules were used and in the second, fixed precision stopping was used. In the third and fourth sets the $\delta_{i2}-\delta_{i1}$ were allowed to vary with the restriction $0<\delta_{i2}-\delta_{i1}<1.5$ placed on the items. The third set used fixed length stopping rules and the fourth, fixed precision.

In each of these sets of simulations the restrictions on the item difficulties ensures that the item information functions are maximized at $\delta_{i\bullet}= (\delta_{i2}-\delta_{i1})/2$ and at this point the expected score of a person with ability $\beta=\delta_{i\bullet}$ is $m_i/2$ (in this case, $m_i/2=1.0$). Consequently the problem of comparing item selection criteria as discussed in the previous chapter is avoided. The design for the four sets of simulations is shown in Table 3.1.

Table 3.1 Design of the simulations

Simulations	$\delta_{i2}-\delta_{i1}=1$	$0<\delta_{i2}-\delta_{i1}<1.5$	Fixed length	Fixed Prec
One	●		●	
Two	●			●
Three		●	●	
Four		●		●

For each of the simulations five item bank sizes were used; 50, 75, 100, 150 and 200 items and for each bank the average difficulties $\delta_{i\bullet}$ were rectangularly distributed from -3 to 3 logits. For the first two sets of simulations $\delta_{i1}=\delta_{i\bullet}-0.5$ and $\delta_{i2}=\delta_{i\bullet}+0.5$ were then calculated while for the third and fourth sets a difference $d_i=\delta_{i2}-\delta_{i1}$ was generated using a uniformly distributed psuedo-random number generator and $\delta_{i1}=\delta_{i\bullet}-d_i/2$ and $\delta_{i2}=\delta_{i\bullet}+d_i/2$ were then calculated. The student ability parameters β_n were rectangulary distributed between -3.5 and +3.5 logits and the population was fixed at 1000 students. Each student's response to an item was simulated by calculating the model probability of each possible response given the fixed item paramaters and the generating value of β_n and then comparing this modelled probability to a uniformly distributed psuedo-random number to determine the response. In the estimation procedure the item difficulty parameters were not calibrated but treated as fixed and used to estimate the abilities.

For the fixed length analysis students were administered tests of length 5, 10, 15, 20, 25 and 30 items for each of the five item banks. That is, every student was given 30 items and ability estimates were recorded after 5, 10, 15, 20, 25 and 30 items. For the fixed precision analyses, four stopping rules were used (i) $se(\beta_n)<0.6$ or $length>10$, (ii) $se(\beta_n)<0.5$ or $length>15$, (iii) $se(\beta_n)<0.4$ or $length>20$ and (iv) $se(\hat{\beta}_n)<0.3$ or $length>30$. Every student was given a maximum of 30 items and ability estimates and standard errors were recorded after each of the four criteria were first satisfied. Abilities were estimated using a Newton-Raphson maximum likelihood procedure and items were selected from the bank so that they maximized the information function. In this case, maximum information is provided by the item with the mean difficulty, $\delta_{i\bullet}$, closest to the current ability estimate.

In each case testing commenced with the administration of the item of median difficulty. After an ability estimate was made the bank was searched and the unused item that provided the most information was selected for administration. Since the maximum likelihood procedure cannot provide an ability estimate for a student with a zero or perfect score, a fixed increment of 0.7 was used until a maximum likelihood estimate could be made (i.e., if a student had a zero score then 0.7 logits were subtracted from that student's current ability estimate; if a student had a perfect score then 0.7 logits were added to the ability estimate). As soon as a student's total test score was no longer perfect or zero a maximum likelihood estimate could be made. Using an increment of 0.7 ensured that if a student scored zero on all items it would take five items to reach the lower extreme of the item bank. At that stage testing was terminated. Similarly if a student achieved a perfect score on each item it would take five items to reach the upper limit of the bank (see Patience and Reckase, 1980). If an estimate could not be made after the administration of five items, testing was terminated and no ability estimate was provided for that student at any test length.

METHODS OF ANALYSIS

The methods of analysis that were employed focus on precision and recovery. Precision is viewed as the accuracy, in terms of standard error, that can be achieved for individuals over the whole ability range. Recovery is viewed as the accuracy with which the adaptive procedure is able to recover the generating values used in the simulations.

Precision

(i) *Information:* The test information (in this case the reciprocal of the square of the asymptotic standard error) provides an index of test precision at all levels of ability. By examining the information curves, the effect of the bank size and test length on the suitability range of the test is determined in terms of measurement precision. For the first two sets of simulations $\delta_{i2}-\delta_{i1}=1$ is fixed so the formula (3.12) allows the calculation of the maximum possible amount of information that can be provided by items of this type if an item with $\delta_{i\bullet} = \hat{\beta}$ is always available.

Item information is defined at the ability estimate, $\hat{\beta}_n$, for student n by (2.6) as:

$$I_i(\beta_n) = \sum_{k-1}^{m_i} k^2 P_{X_i}(\beta_n) - (\sum_{k-1}^{m_i} k P_{X_i}(\beta_n))^2$$

and test information as:

$$I(\beta_n) - \sum_{i-1}^{L} I_i(\beta_n)$$

As discussed in the previous chapter the standard error of an ability estimate, $\hat{\beta}$, corresponds to the square root of the reciprocal of the test information. That is:

$$se(\beta) = 1/\sqrt{I}(\beta)$$

Recovery

(ii) *Bias:*The possibility of any consistent under or over estimation in the abilities was studied by calculating the average bias. The total population of 1000 was broken into succesive subgroups of 100 according to their generating ability and then for each group the average difference between the generating and recovered abilities was calculated.

The average bias for each group of 100 students is given by:

$$B = \frac{\sum(\hat{\beta}_n - \beta_n)}{100}$$

(iii) *Efficiency:*The efficiency of the procedure in recovering abilities was examined by comparing the root mean square:

$$RMS = \left[\frac{\sum(\hat{\beta}_n - \beta_n)^2}{100} \right]^{1/2}$$

with the average standard error:

$$ASE = \left[\frac{\sum se(\beta_n)^2}{100} \right]^{1/2}$$

Studying efficiency in this way indicates whether the observed variance in ability estimates is the same as the modelled variance. This approach also gives a check for the asymptotic standard errors produced by the maximum likelihood estimation procedure. Since, if no bias is evident, the RMS is an empirically based estimate of the standard error for data that is generated to fit the model. This is of particular interest for the shorter tests were the asymptotic result is less likely to hold.

(iv) *Correlations between Generating and Estimated Abilities:* Estimated abilities were correlated with the generating values of the abilities, and the correlation coefficients analysed. This method is widely used in the examination of the recovery power of simulations. These correlations are often called fidelity or validity correlations (eg., Weiss, 1982)

Table 3.2 Theoretical Limits on Information and Standard Error

Test Length	Max Information	Min Standard Error
5	2.74	0.60
10	5.48	0.43
15	8.22	0.35
20	10.96	0.30
25	13.70	0.27
30	16.44	0.25

PRECISION

Information

Simulation Set One: Recall that simulation set one used banks of items with δ_{i2}-δ_{i1}=1 and fixed test length stopping criteria. Since δ_{i2}-δ_{i1}=1 for all items, applying formula (2.12) it is possible to calculate the theoretical limits on the test information. This is the amount of information that could be provided if an unlimited supply of items were available ensuring that β=$\delta_{i\bullet}$ was satisified for all items selected for administration. Table 3.2 shows the upper limits for information and the corresponding lower limits of the standard error for tests of lengths: 5, 10, 15, 20, 25, and 30 items, where δ_{i2}-δ_{i1}=1.

Figure 3.1 shows plots of the test information curves for each of the six test lengths when using the 50 item bank. The six horizontal bars drawn in Figure 3.1 (and Figures 3.2, 3.3 and 3.4) indicate the maximum information that can be provided by these tests as shown in Table 3.2. The general shape of the curves indicates floor and ceiling effects in the item bank. At the upper and lower extremes of the ability range there are fewer appropriate items and this leads to a decrease in precision. At this bank size, it is apparent that little is gained beyond 20 items in terms of test information and consequently measurement precision. At the centre of the ability distribution the increase from 20 to 30 items increases information from about 8.5 to 12.5, which is equivalent to a decrease in the standard error from 0.34 to 0.28. At the extremes of the ability distribution, where the available supply of suitable items is smaller, each additional item beyond about 15 items did not provide any substantial increase in precision.

Beyond about 15-20 items students of high ability will be administered items that they find very easy because the items of appropriate difficulty have been exhausted. Similarly at the low ability end only items that have too high a difficulty will be available for administration.

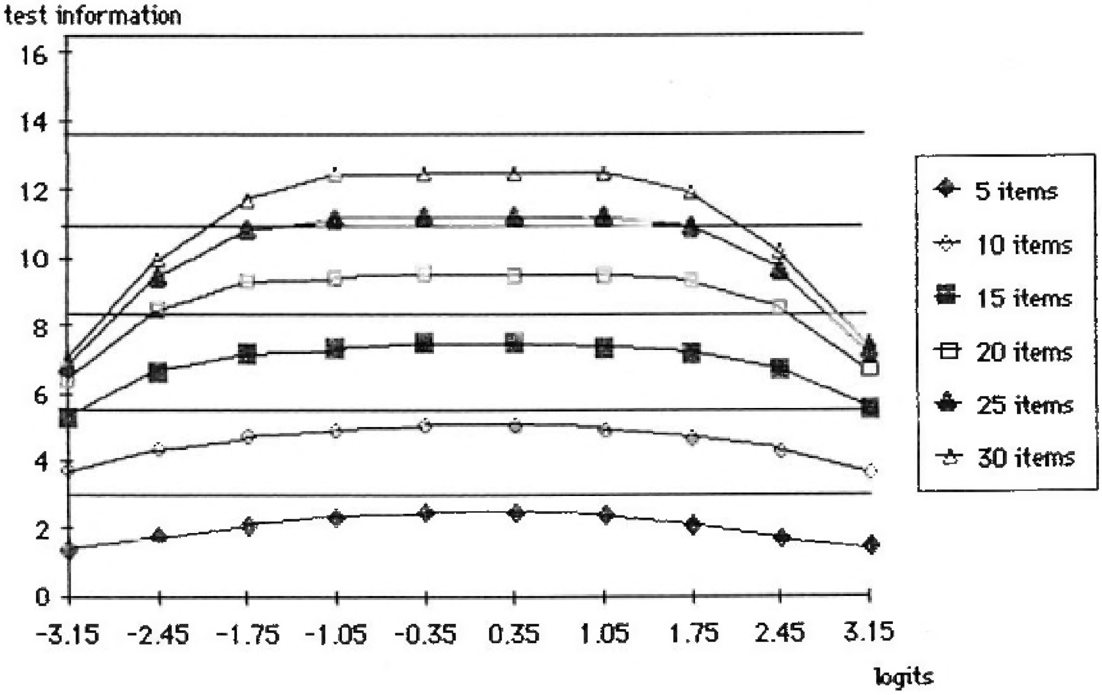

Figure 3.1 Information curves for tests drawn from a 50 item bank

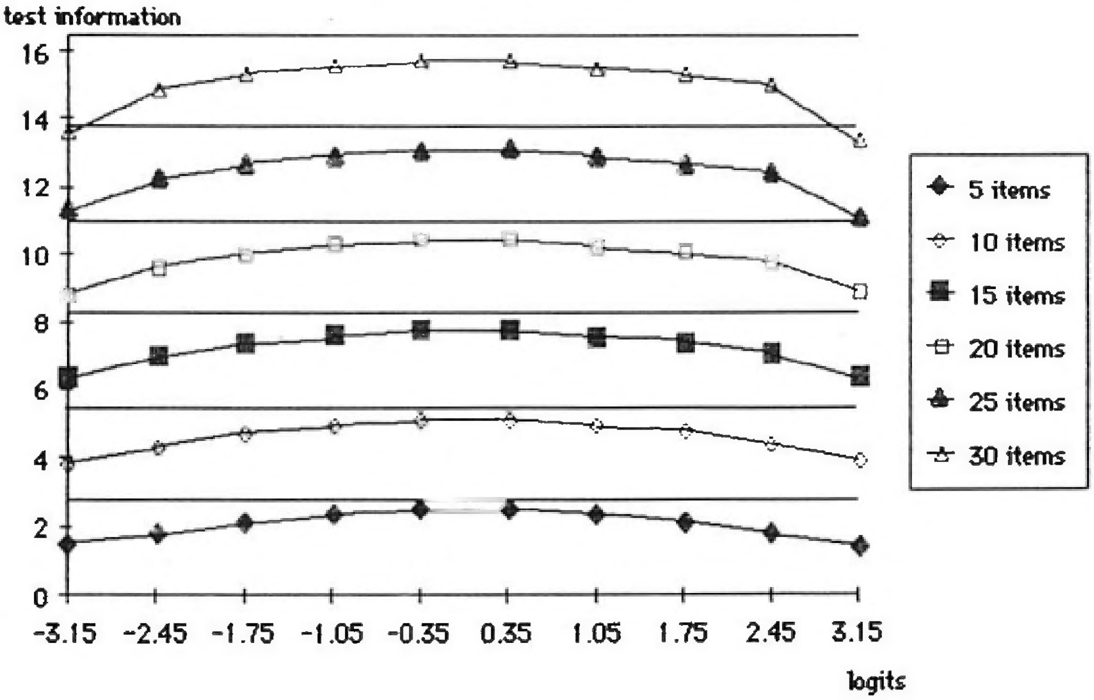

Figure 3.2 Information curves for tests drawn from a 200 item bank

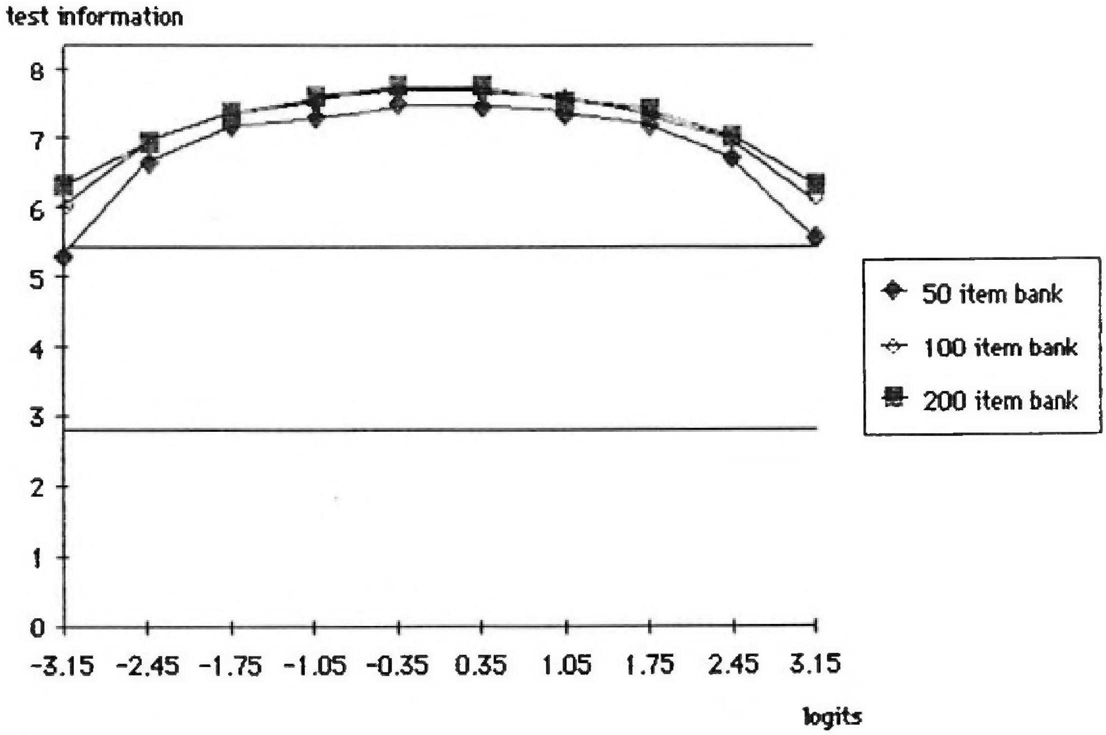

Figure 3.3 Information curves for 15-item tests drawn from three different item bank sizes

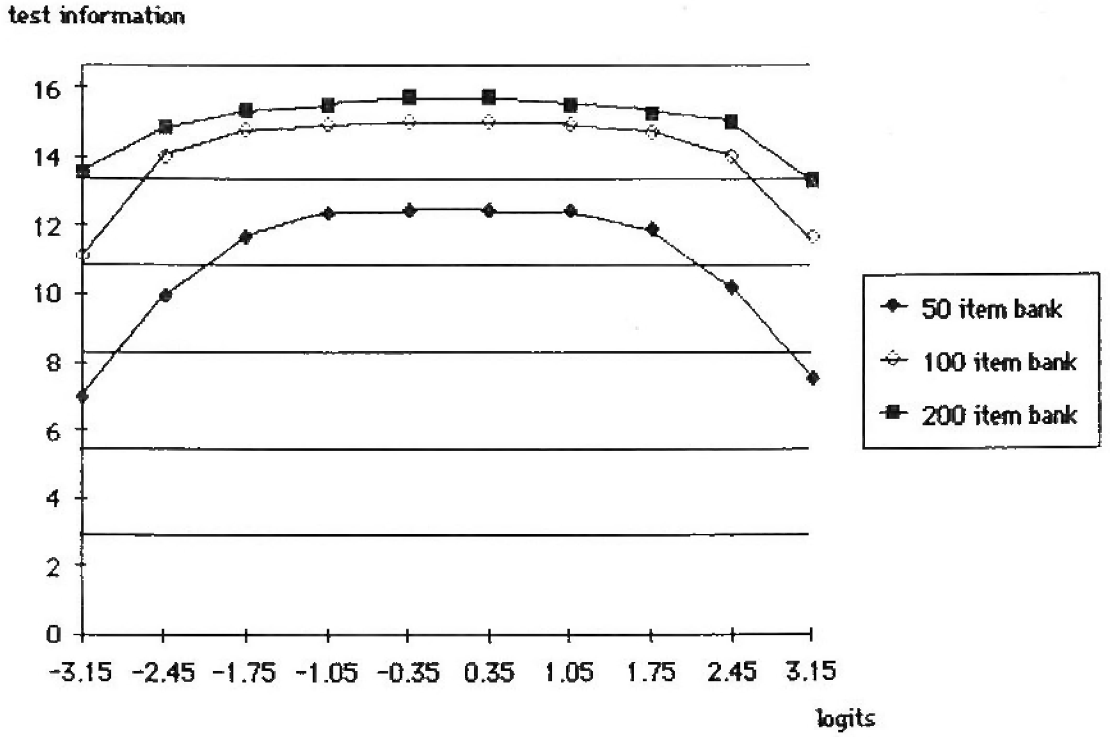

Figure 3.4 Information curves for 30-item tests drawn from three different item bank sizes

Figure 3.2 shows the corresponding results for the 200 item bank. In this case the increase in test length beyond 20 items continued to contribute to the total information because of the availability of more appropriate (informative) items. This is indicated by the information curves reaching the horizontal maximum information lines for each test length. Note however, that the increase in information from 20 to 30 items only corresponds to a decrease in the standard error from about 0.30 to 0.25 at the middle of the ability distribution, and from about 0.33 to 0.28 at the extremes.

Figure 3.3 shows the test information curves again, this time for the 15-item tests using three different item bank sizes and Figure 3.4 shows the corresponding curves for the 30-item tests. In Figure 3.3 we see almost no improvement in measurement precision by extending the item bank beyond 50 items. However, Figure 3.4 shows a larger item bank can improve accuracy if longer tests are required, but the gains are not substantial. The largest gains are at the extremes of the ability distribution. These results clearly fit with the floor and ceiling effects described above. For the 15-item tests a bank of 50 items is sufficient to ensure the administration of appropriate items, so little improvement in precision is gained by an increase in bank size. For the 30 item test the larger item bank increases precision, particulary at the extremes because the number of appropriate items is limited when trying to select 30 items from the smaller item bank.

All bank sizes show floor and ceiling effects that result in a drop in precision at the extremes of the ability distribution. This is most marked for large tests drawn from smaller item banks. As a result, larger item banks may be necessary for longer tests. However, the amount of improvement gained by administering more than 15-20 items is small even when using banks of 200 items.

Simulation Set Two: In simulation Set Two the items are the same as those in set one, but rather than administering fixed length tests, stopping criteria that specify a fixed level of precision have been used. Table 3.3 shows the theoretical minimum number of items (using (2.12)) required to reach each of the prespecified levels of precision used as stopping rules. In Figures 3.5 and 3.6 the horizontal lines indicate the minimum required test lengths that are shown in Table 3.3.

Table 3.3 Theoretical Minimum Test Lengths required to achieve specified levels of precision

Specified Standard Error	Information	Minimum Test Length
0.6	2.78	5.07
0.5	4.00	7.30
0.4	6.25	11.41
0.3	11.11	20.28

Figure 3.5 shows the average test lengths for the four fixed precision stopping rules when drawn from a 50 item bank. For all levels of precision a slightly greater number of items has been required to establish the same precision at the extremes of the ability distribution. For the se(β_n)<0.6 and se(β_n)<0.5 criteria the precision requirements were met before the test length limits had been exceeded. For the third level of precision (se(β_n)<0.4) however floor and ceiling effects began to emerge at the extremes as the precision requirements could not always be met before the maximum test length was exceeded. Since the average test length is less than the fixed limit of 20 items a number of students satisfied the precision requirement but a large number must have received the maximum number of items without acceptable standard errors. For the strongest criterion (se(β_n)<0.3) the bank size was sufficient for the middle of the ability distribution but it was unable to ensure equi-precise measurement for students at the extremes. This is illustrated by the flattening of the average test length at the tails of the se(β_n)<0.3 curve.

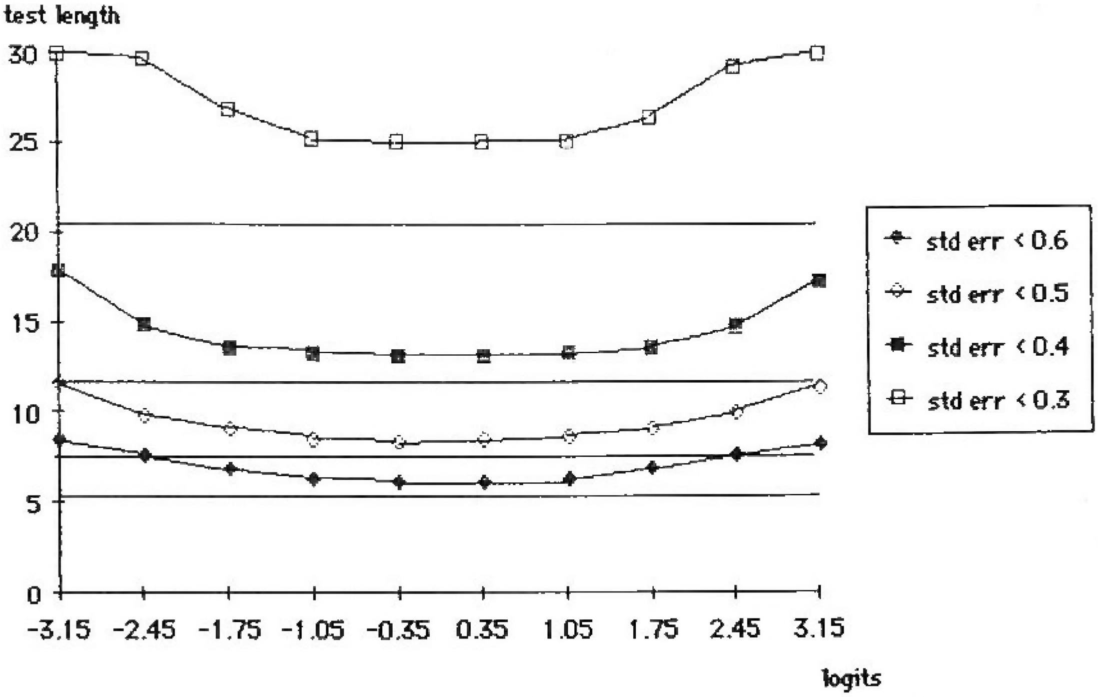

Figure 3.5 Test lengths for tests drawn from a 50 item bank

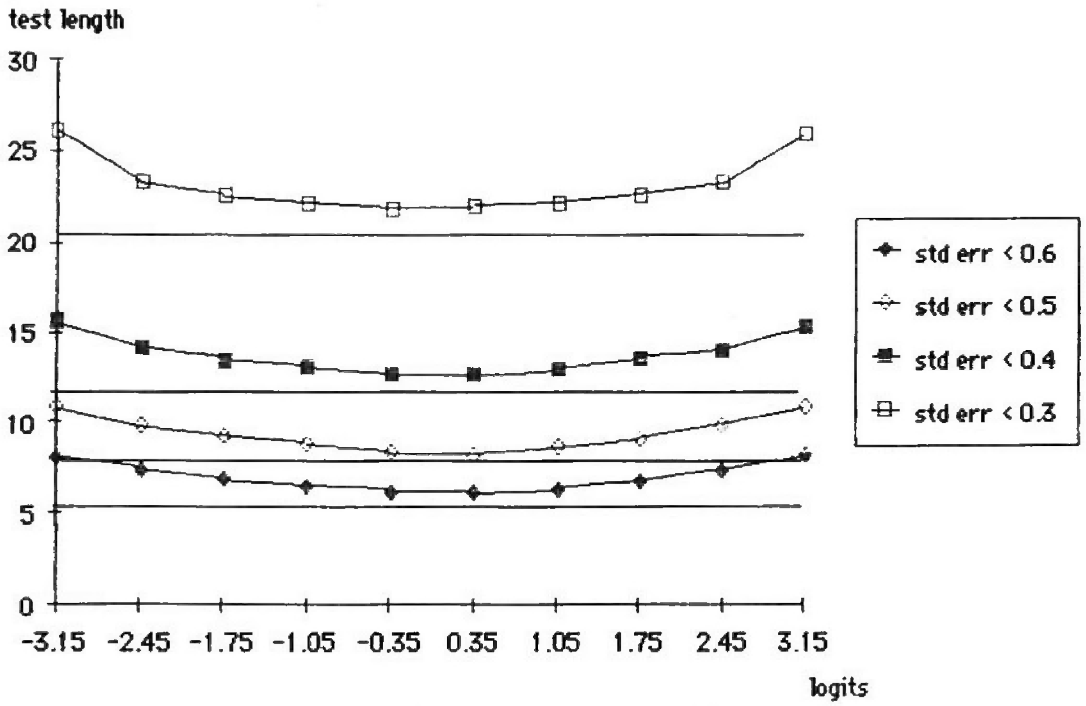

Figure 3.6 Test lengths for items drawn from 150 item bank

Figure 3.6 shows the average test lengths for tests drawn from a 150 item bank. In this case all but the most precise tests achieved equi-precise measurement over the ability range of interest. For the se(β_n)<0.3 criterion some students at the extremes of the ability distribution were not measured with the required accuracy. For the three shorter tests two to three more items are required at the extremes to ensure equi-precise measurement.

A comparison of the average test lengths in Figures 3.5 and 3.6 indicates that for all but the longest test the smaller item bank can produce estimates of equal precision with about the same number of items. For the se(β_n)<0.6 criterion the lengths are almost identical, for the se(β_n)<0.5 criterion there is an advantage of about one item at the extremes and for the se(β_n)<0.4 criterion there is an advantage of about one item except at the extremes were the 50 item bank showed stronger floor and ceiling effects.

Table 3.4 Theoretical limits on information and standard error

Test Length	Max Information	Min Standard Error
5	2.89	0.59
10	5.79	0.42
15	8.67	0.34
20	11.58	0.29
25	14.45	0.26
30	17.34	0.24

Simulation Set Three: The next two sets of simulations are for the partial credit model because they allow δ_{i2}-δ_{i1} to vary, within the constraints $0<\delta_{i2}-\delta_{i1}<1.5$. The between item variation means that it is not possible to establish an exact theoretical framework for information and test length using (2.12) but if an average of $\delta_{i2}-\delta_{i1}=0.75$ is assumed, then the maximum information and minimum standard errors in Table 3.4 can be used as a guide.

Table 3.4 indicates that we can expect slightly improved precision for this item bank than for the item bank used in simulation set one and two. This is because the smaller average value of δ_{i2}-δ_{i1} corresponds to a more peaked information curves.

Figures 3.7 and 3.8 show the information curves for six tests drawn from the item banks with variable δ_{i2}-δ_{i1}. As expected the curves are almost identical to those shown in Figures 3.1, 3.2, 3.3, and 3.4. Floor and ceiling effects are evidenced in all tests by the drop in information at the extremes of the ability distribution. This is particularly true for the longer tests drawn from the smaller item banks. As with previous results it would appear that for tests of up to about 15 items the smaller item banks may be sufficient.

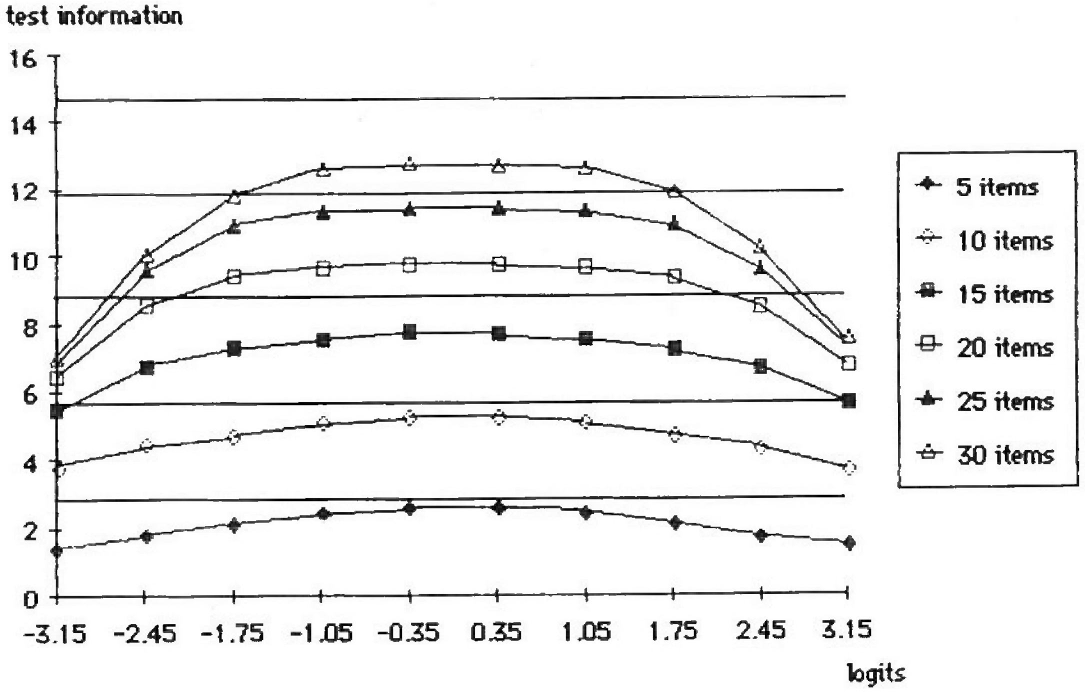

Figure 3.7 Information curves for tests drawn from a 50 item bank

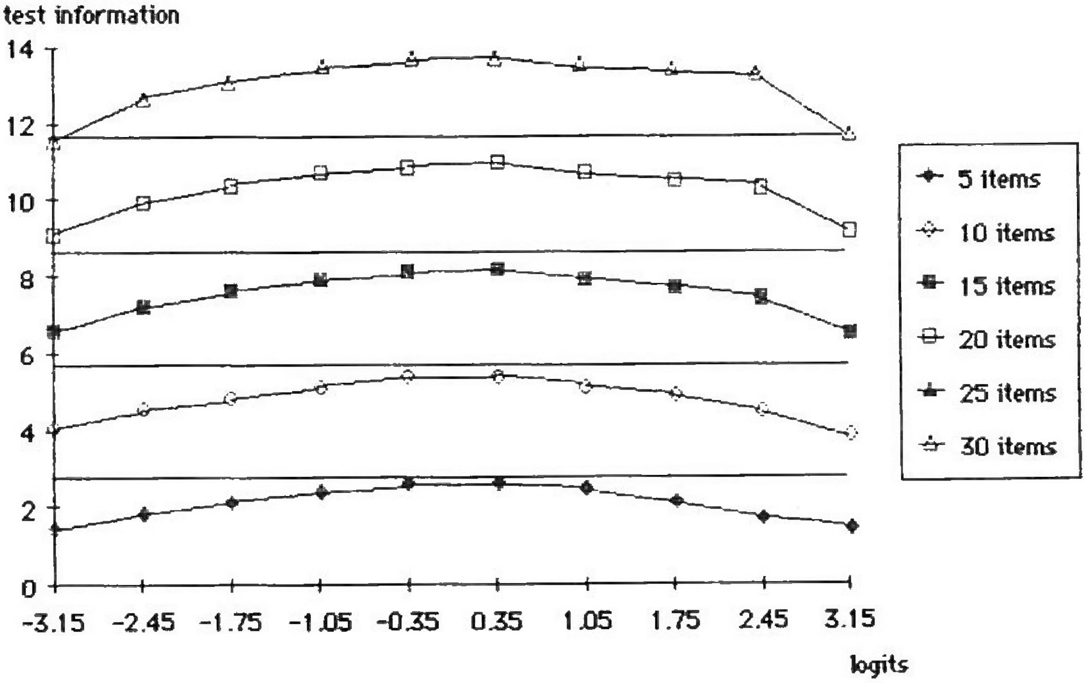

Figure 3.8 Information curves for tests drawn form 200 item bank

A close comparison of Figure 3.1 and Figure 3.7 shows that the information provided by the tests drawn from the rating scale banks is slightly less than that drawn from the partial credit banks. This is due to the higher average maximum item information provided by items in the variable difference bank. As shown in the above discussion about information, as $\delta_{i2}-\delta_{i1}$ decreases, the information functions become more peaked and have a greater maximum. Since the differences $\delta_{i2}-\delta_{i1}$ in the variable difference bank are rectangulary distributed between 0 and 1.5 the average maximum item information is slightly greater than the bank where $\delta_{i2}-\delta_{i1}$ is fixed at one.

Simulation Set Four: In this simulation the fixed precision stopping criteria are used. By taking an average difference of $\delta_{i2}-\delta_{i1}=0.75$ we use (2.12) to produce Table 3.5 which gives the minimum test lengths that would be required to reach the prescribed levels of precision. A comparison of Table 3.3 shows how the greater peak in the information curves for these items leads to fewer items being required to reach the prescribed levels of precision.

Table 3.5 Theoretical minimum test lengths required to achieve specified levels of precision

Specified Standard Error	Information	Minimum Test Length
0.6	2.78	3.80
0.5	4.00	6.91
0.4	6.25	10.80
0.3	11.11	19.19

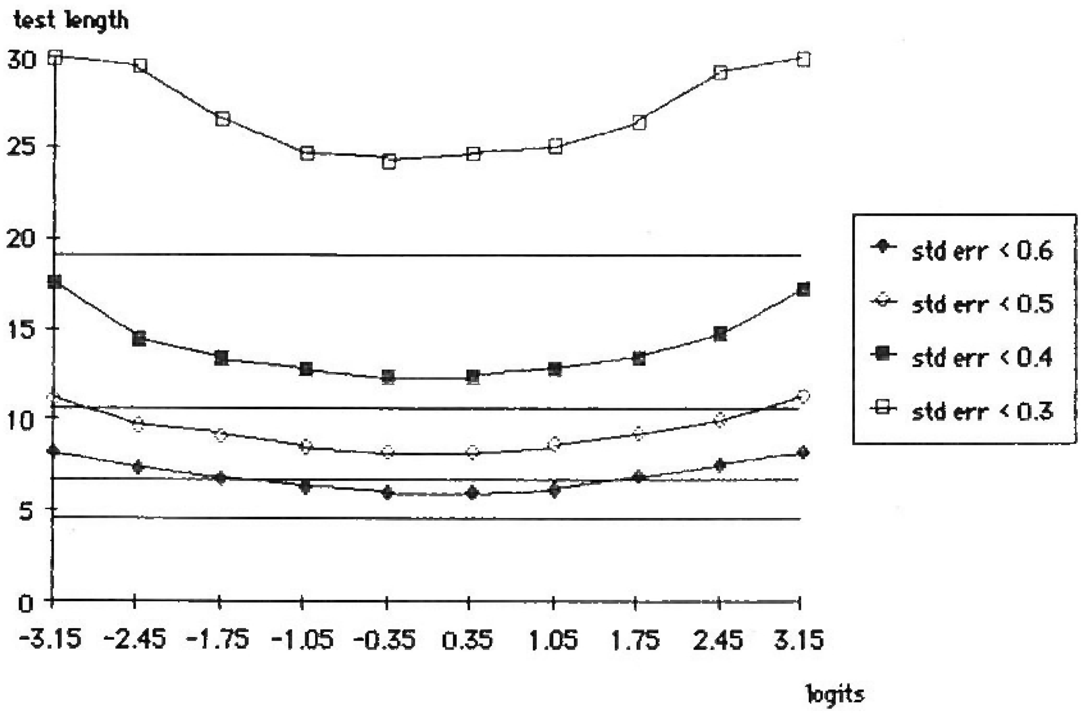

Figure 3.9 Test length for tests drawn from a 50 item bank

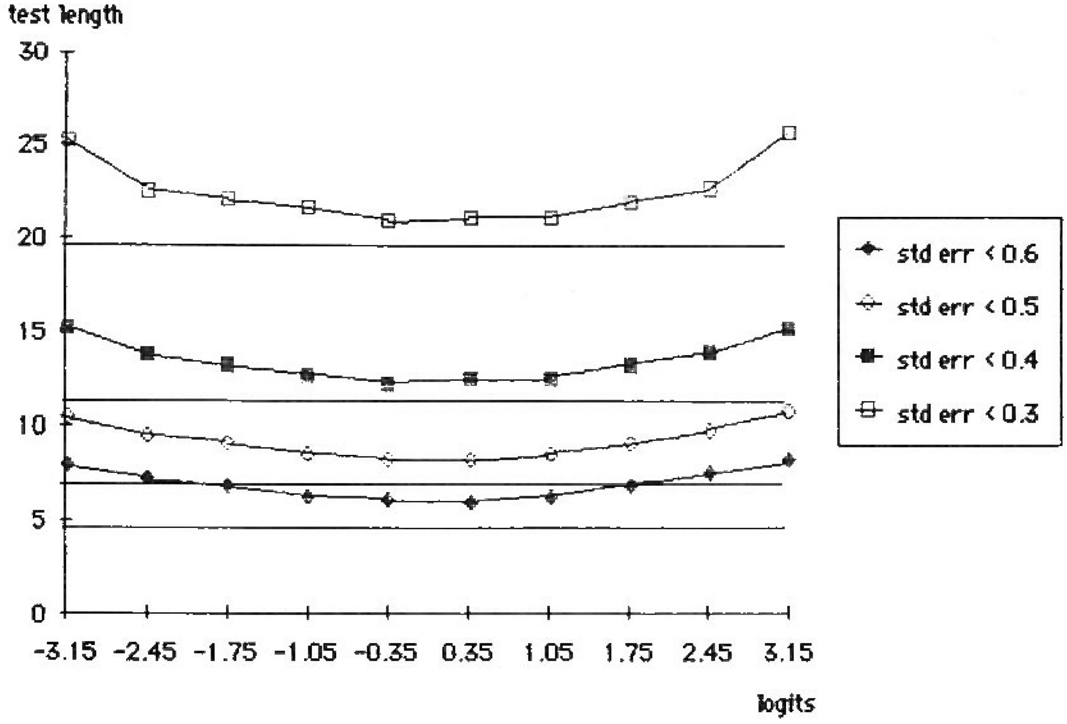

Figure 3.10 Test length for tests drawn from a 150 item bank

The results in Figure 3.9 and 3.10 are almost identical to those shown in Figure 3.5 and 3.6. As we would expect from the comparison of Tables 3.4 and 3.6 there is a slight advantage in favour of the variable differences in terms of test length. As with the improvement mentioned above this is due to the higher average information provided by each item in the bank.

RECOVERY

Bias

Tables 3.6, 3.7, 3.8, and 3.9 show the average differences between the recovered and generating abilities for 10 groups of 100 students grouped according to their ability. Positive values indicate that the recovered abilities are greater than the generating abilities, while negative values indicate that the recovered abilities are less than the generating abilities.

Tables 3.6, 3.7, 3.8 and 3.9 indicate three trends; (1) as the test length increases there is greater stability since the bias values become smaller, (2) as the bank size increases there is some tendency for greater stability and (3) there is greater stability at the middle of the ability distribution where fluctuations about zero appear to be smaller.

Overall there does not appear to be any indication of bias in the ability estimation. The values in the tables are consistently small and would appear to result form random fluctuations.

Table 3.6 Average differences between generating and recovered abilities for fixed length, fixed difference simulations

	Bank Size								
	50 Items			100 items			200 items		
	Test Length			Test Length			Test Length		
ability	5	20	30	5	20	30	5	20	30
-3.15	.13	.00	-.04	.02	-.04	-.05	.22	.02	.04
-2.45	-.02	.02	.01	.13	.05	.06	.05	-.01	.00
-1.75	-.07	.00	-.02	-.08	.02	.00	-.04	-.03	-.02
-1.05	-.03	.00	.00	-.19	-.08	-.04	-.01	.03	.00
-0.35	-.08	.01	.02	-.05	.01	.02	.02	-.01	-.02
0.35	-.06	.00	.01	.05	.00	-.01	-.03	.01	.03
1.05	-.05	-.02	.00	.04	.01	-.01	-.02	.00	.00
1.75	.03	.03	-.03	.05	.00	.02	-.02	-.06	-.02
2.45	.13	.01	.00	.11	.00	.01	.04	-.01	.01
3.15	-.09	.04	.05	-.18	-.03	-.02	.00	.00	. 03

Table 3.7 Average differences between generating and recovered abilities for variable length, fixed difference simulations

	Bank Size											
	50 Items				100 items				200 items			
	Test Length				Test Length				Test Length			
ability	1	2	3	4	1	2	3	4	1	2	3	4
-3.15	-.13	-.06	.00	.04	-.12	-.05	-.03	-.03	-.20	-.12	-.07	-.06
-2.45	-.06	-.08	-.04	-.01	-.17	-.08	-.05	-.07	-.11	-.02	-.02	.01
-1.75	-.01	-.04	.01	.01	-.01	.00	-.04	-.02	-.05	-.03	.00	.03
-1.05	-.04	-.09	-.04	-.01	.16	.17	.11	.06	-.05	-.05	-.04	-.02
-0.35	.04	-.02	.01	-.01	-.02	.01	.01	.00	.00	.01	-.02	.00
0.35	.08	.09	.04	.00	-.04	-.02	.00	.01	.00	.00	.00	.00
1.05	.07	.03	.03	.01	-.01	-.02	-.03	.00	.07	.10	.02	.00
1.75	.00	.00	.00	-.02	.03	.02	.02	.00	.09	.13	.07	.06
2.45	-.06	-.04	.01	.00	-.02	-.04	-.01	.02	.04	. 00	.03	.01
3.15	.09	.04	-.01	-.05	.22	.14	.05	.06	.04	.05	.02	.00

Table 3.8 Average differences between generating and recovered abilities for fixed length, variable difference simulations

	Bank Size								
	50 Items			100 items			200 items		
	Test Length			Test Length			Test Length		
ability	5	20	30	5	20	30	5	20	30
-3.15	-.02	-.03	-.01	.00	-.04	-.04	.14	.06	.05
-2.45	-.03	.01	-.02	.09	.06	.05	.03	-.03	.00
-1.75	.14	.02	.01	-.05	-.01	.00	.09	.00	.00
-1.05	-.11	.00	.01	-.02	-.01	-.02	-.09	.00	.00
-0.35	-.05	-.01	-.01	.03	.01	.02	-.08	-.01	-.02
0.35	-.03	-.03	-.03	.13	.01	-.01	.05	.02	.00
1.05	-.05	-.02	-.01	.07	-.02	-.02	.01	-.02	-.01
1.75	-.06	-.04	-.03	.01	.02	.03	.15	.01	.00
2.45	-.07	.02	.01	.11	-.03	.01	.08	.00	.01
3.15	-.06	-.01	.01	-.16	-.03	-.03	-.14	-.01	.01

Table 3.9 Average differences between generating and recovered abilities for variable length, variable difference simulations

	Bank Size											
	50 Items				100 items				200 items			
	Test Length				Test Length				Test Length			
ability	1	2	3	4	1	2	3	4	1	2	3	4
-3.15	.19	.09	.02	-.05	.15	.09	.03	-.03	.26	.17	.12	.08
-2.45	.09	.09	.06	.03	.19	.11	.09	.07	.17	.11	.01	.01
-1.75	.00	-.01	-.01	.01	.03	.03	.02	-.01	.03	-.05	-.01	-.01
-1.05	-.01	.04	.03	.00	-.04	-.05	-.02	.00	-.04	-.02	-.01	.00
-0.35	-.05	-.03	-.02	.02	.04	.01	.01	.01	.00	.01	-.03	-.01
0.35	-.14	-.06	-.07	.00	.09	.04	.03	.01	.02	.02	.02	.02
1.05	-.07	-.03	-.04	.00	.04	.02	.01	-.02	-.04	-.02	-.03	-.02
1.75	-.01	-.04	-.03	.02	-.01	.05	.04	.03	.02	.05	-.02	.01
2.45	-.03	-.04	.04	-.01	.01	.01	.00	-.02	-.01	-.01	-.01	-.01
3.15	-.16	-.09	-.04	.04	-.21	-.11	-.06	-.07	-.06	-.04	-.03	-.02

Table 3.10 Root mean square and average standard error for simulation set one

Test	Item Bank Size				
Length	50	75	100	150	200
RMSE					
5	.67	.68	.72	.71	.71
10	.44	.47	.48	.46	.46
15	.38	.37	.39	.37	.38
20	.34	.33	.33	.31	.32
25	.31	.30	.30	.27	.28
30	.31	.28	.27	.25	.25
ASE					
5	.74	.73	.74	.74	.74
10	.47	.47	.47	.47	.47
15	.39	.38	.38	.37	.37
20	.35	.33	.32	.32	.32
25	.32	.30	.30	.29	.28
30	.31	.28	.27	.26	.26
Ratio RMSE/ASE					
5	.92	.93	.97	.97	.96
10	.94	.99	1.02	.97	.98
15	.98	.99	1.04	.99	1.01
20	.99	.99	1.03	.97	.99
25	.97	1.00	1.02	.96	.98
30	.98	1.01	1.01	.97	.98

Table 3.11 Root mean square and average standard error for simulation set two

Stopping	Item Bank Size				
Criteria	50	75	100	150	200
RMSE					
se(β)<0.6	.55	.55	.58	.56	.56
se(β)<0.5	.46	.47	.49	.46	.47
se(β)<0.4	.38	.39	.41	.38	.39
se(β)<0.3	.31	.30	.30	.29	.30
ASE					
se(β)<0.6	.58	.58	.58	.57	.57
se(β)<0.5	.49	.49	.49	.48	.48
se($\hat{\beta}$)<0.4	.40	.39	.39	.39	.39
se(β)<0.3	.32	.30	.30	.30	.30
Ratio RMSE/ASE					
se(β)<0.6	.96	.96	1.01	.97	.98
se(β)<0.5	.96	.97	1.01	.96	.98
se(β)<0.4	.96	.98	1.04	.97	.99
se(β)<0.3	.98	.99	1.01	.97	1.00

Table 3.12 Root mean square and average standard error for simulation set three

Test	Item Bank Size				
Length	50	75	100	150	200
RMSE					
5	.73	.73	.73	.73	.73
10	.47	.47	.47	.46	.46
15	.38	.37	.37	.37	.37
20	.34	.32	.32	.31	.31
25	.32	.30	.29	.28	.28
30	.31	.28	.26	.26	.26
ASE					
5	.68	.68	.68	.69	.70
10	.44	.48	.45	.45	.45
15	.38	.38	.37	.36	.37
20	.33	.32	.32	.31	.30
25	.31	.30	.28	.27	.27
30	.30	.28	.26	.25	.24
Ratio RMSE/ASE					
5	.93	.93	.93	.95	.96
10	.94	1.02	.98	.98	.98
15	1.00	1.03	1.04	.97	1.00
20	.97	1.00	.97	1.00	.97
25	.97	1.00	.97	.96	.98
30	.97	1.00	1.00	.96	.96

Table 3.13 Root mean square and average standard error for simulation set four

Stopping	Item Bank Size				
Criteria	50	75	100	150	200
RMSE					
se(β)<0.6	.54	.55	.58	.56	.56
se(β)<0.5	.46	.47	.49	.46	.47
se(β)<0.4	.38	.39	.41	.38	.39
se(β)<0.3	.31	.30	.30	.29	.30
ASE					
se(β)<0.6	.58	.58	.58	.57	.57
se(β)<0.5	.49	.49	.49	.48	.48
se(β)<0.4	.40	.39	.39	.39	.39
se(β)<0.3	.32	.30	.30	.30	.30
Ratio RMSE/ASE					
se(β)<0.6	.96	.96	1.01	.97	.98
se(β)<0.5	.96	.97	1.01	.96	.98
se(β)<0.4	.96	.98	1.04	.97	.99
se(β)<0.3	.98	.99	1.01	.97	1.00

Table 3.14 Correlations between recovered and generating abilities

	ITEM BANK SIZE				
Stopping criteria	50	75	100	150	200
Set 1					
5	.95	.94	.94	.94	.94
10	.98	.97	.97	.97	.97
15	.98	.98	.98	.98	.98
20	.99	.99	.99	.99	.99
25	.99	.99	.99	.99	.99
30	.99	.99	.99	.99	.99
N	956	967	967	961	962
Set 2					
1	.97	.96	.96	.96	.96
2	.97	.97	.97	.97	.97
3	.98	.98	.98	.98	.98
4	.99	.99	.99	.99	.99
N	955	967	968	961	962
Set 3					
5	.94	.94	.95	.94	.94
10	.98	.97	.98	.98	.98
15	.98	.98	.98	.98	.98
20	.99	.99	.99	.99	.99
25	.99	.99	.99	.99	.99
30	.99	.99	.99	.99	.99
N	956	968	969	963	962
Set 4					
1	.96	.96	.96	.96	.96
2	.97	.97	.97	.97	.97
3	.98	.98	.98	.98	.98
4	.99	.99	.99	.99	.99
N	960	964	966	956	956

Efficiency

Tables 3.10, 3.11, 3.12, and 3.13 show the average root mean square, the average standard error and the ratios of the average root mean square to the average standard error. In examining these tables it should be noted that the values in a column are not independent. Since the ability estimates were recorded after the administration of 5, 10, 15, 20, 25 and 30 items for the fixed length analyses, and after the stopping rules were first satisfied for the fixed precision analyses the ability estimates are based on a number of items that were also included in the previous ability estimate. These tables illustrate a number of trends. Firstly, as test length increases both the root mean square and the average standard error decrease. Secondly, as bank size increases, both the root mean square and the average standard errors decrease. These results were expected since longer tests are more accurate -- if the items are appropriate -- and if a larger item bank is used the appropriateness of the items administered to an individual student should increase.

The ratios RMSE/ASE compare the observed variation in the ability estimates with the modelled variation. If the adaptive testing procedure is working accurately the root mean square values should be equal to the average standard error and consequently the ratios should be equal to one. The nearness of the ratios to unity along with evidence of no bias indicates good recovery.

The nearness of the ratios to unity also provides some evidence to support the use of asymptotic standard errors even when short tests are being used. There is some evidence that the ratios are closer to unity for the longest of the tests but the improvement is slight.

Correlations

Table 3.14 shows the correlations between the generating and recovered values of the abilities along with the number of abilities that were succesfully estimated by each test.

These correlations are all very close to one and suggest that little is to be gained by administering more than about 15 items or by using a bank of more than 50 items when all items are scored in three ordered categories and data conform to the partial credit model. Over this ability range, similar numbers of students could be measured at each bank size.

4. SUMMARY AND CONCLUSIONS

In this report we have begun the exploration of adaptive testing strategies based on the Rasch rating scale and partial credit models. Adaptive testing technology, if applied with these models, could be applied to a wider range of item and test types than is currently possible. This could include attitude and personality scales scored in the likert tradition, problem solving tasks that use scoring that includes credit for partial understanding or partial completion, item clusters and interactive items.

SIMULATIONS

The results of the simulations reported in Chapter 3 indicate the potential of an adaptive testing procedure when items behave according to the model. The evidence regarding optimum test length and bank size provides a useful framework for practical test construction.

The results using fixed length stopping rules indicate good recovery of students' abilities when tests of 5 to 30 items are constructed from banks of 50 to 200 items. Similarly the fixed precision analyses show that tests of 10 to 20 items can provide equi-precise measurement at acceptable levels of measurement precision, even when item banks with as few as 50 items are used.

There appears to be little difference between item banks constructed with $\delta_{i2}-\delta_{i1}$ fixed at one or banks constructed of items with $\delta_{i2}-\delta_{i1}$ varying between 0 and 1.5. A range that is usually considered desirable from a substantive perspective.

The correlations show that the point estimates of students' abilities correlate satisfactorily with generating values of those parameters.

The information curves and the average standard errors show that tests of 15-20 items drawn from item banks of about 50-100 items provide ability estimates with reasonable precision. They also indicate that when a smaller item bank is used there may be little gain in administering more than about 15 items, particularly at the extremes of the ability distribution. This is because, for extreme abilities, there are fewer suitable items in the smaller bank. Any further items that are administered are of inappropriate difficulty. It would probably be desirable to add a condition that terminates testing if items a re not available to add some given amount of additional information.

The average standard errors and the information curve plots show that little is to be gained in terms of the measurement precision beyond tests of about 15-20 items based on item banks with as few as 50-100 items. The ratios of the root mean square to the average standard errors provided by the maximum likelihood procedure are very accurate, even for short tests. With the added evidence from the bias plots it can be concluded that the standard deviation about β as given by the root mean square is approximated closely by the standard errors.

These results compare favourably with those that have been reported for adaptive tests based on dichotomously scored items. Research based on dichotomous items has generally concluded that about 20 items are sufficient when selected from banks of about 120 items, provided the item difficulty distribution has no major gaps. These simulations show that the use of good partial credit items may lead to further improvement in testing efficiency.

INFORMATION FUNCTIONS

A major section of the report focused on the information function for the rating scale and partial credit models. While information functions have played a major role in adaptive testing with latent trait models that allow dichotomous scoring only, the relatively simple nature of these functions has made their application straightforward. For the rating scale and partial credit models the information functions are more complex and their application in adaptive testing is likely to be far from straightforward. The analysis undertaken in chapter 2 has identified a number of important points in relation to these models. It highlights that Rasch models, as a general rule, do not require equal item discrimination. What Rasch models require is, equal item information. When applying the Rasch model the test constructor defines the amount of information provided by the item when the number of possible outcomes are specified. The distribution of the available information over the variable of interest is then determined by the item parameters. Some of the implications of different item information distributions on measurement precision are indicated in the simulations.

For the simulations two different item bank types were used. The first, a rating scale bank, used $\delta_{i2}-\delta_{i1}=1$ for all items (i.e., $\tau_1=-.5$, $\tau_2=.5$) and the second, a partial credit bank, used variable $\delta_{i2}-\delta_{i1}$ with the constraint $0\leq\delta_{i2}-\delta_{i1}\leq1.5$. In the second bank, the difference $\delta_{i2}-\delta_{i1}$ was uniformly distributed between 0 and 1.5, so on average the items in the partial credit bank have more peaked information curves. However, since all items had the same number of categories the total information in the two item banks was equal. The results of the simulations indicate small but consistent advantages in precision and efficiency for tests drawn from the partial credit bank. It is likely therefore that a bank of items with $\delta_{i2}-\delta_{i1}<0$ will lead to even further increases in precision and efficiency. Similarly in the case of banks of items with more than three response categories it is likely that precision and efficiency will be maximized if items with more peaked information functions are included in the bank.

Obviously, further investigation needs to be undertaken to explain these issues more fully, but at a first look it appears that there is some kind of incompatibility between these observations and what we might call the substantive requirements of good measurement.

FUTURE RESEARCH

This report points directly to a range of further research questions. The simulations, for example, did not go beyond items with three response categories. What is gained by using items with perhaps five response categories? Five categories are commonly used in attitude rating scales scored in the likert tradition and adaptive attitude testing is an obvious future application for the RSM. What are the implications of using banks with items scored in different numbers of response categories? Will the items with fewer categories be of any value to measurement or will they be passed over for items with more categories that provide more information? The PCM simulations in chapter 3 varied the difference δ_{i2}-δ_{i1} in a limited way. What happens if δ_{i2}-δ_{i1} is allowed more freedom to vary? For items with more than three response categories many variations on the relationships between the δ_{i1}, δ_{i2} .. δ_{im} are possible. Other possibilities include -- What is the effect of different distributions of item parameters in the bank? And how do various estimation methods compare

REFERENCES

Andrich, D.A. (1978). A rating formulation for ordered response categories. Psychometrika 43, 561-573.

Bejar, I.I. (1976). Applications of adaptive testing in measuring achievement and performance. mimeo. (ED 169066)

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees ability. In Lord, F.M., & Novick, M.R. Statistical Theories of Mental Test Scores. Reading, MA: Addison Wesley.

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

Dodd, B.G., and Koch, W.R. (1985) Item and scale information functions for the partial credit model. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago

Lord, F.M. (1980). Application of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika. 47, 147-174.

Masters, G.N. (1984). Constructing and item bank using partial credit scoring. Journal of Educational Measurement 21, 19-32.

Masters, G.N. and Evans, J. (1986) Banking Non-dichotomously scored items. Applied Pshcyological Measurement 10(4) 355-367

Masters, G.N. and Wright, B.D. (1984). The essential process in a family of measurement models. Psychometrika. 49(4), 529-544

McBride, J.R. (1979). Mental Testing: The state of the art. (Research Report ARI-RR-423). Alexandria, V.A.: Army Research Institute for the Behavioral and Social Sciences. (ED 200612)

McBride, J.R. (1976). Simulation studies of adaptive testing: A comparative evaluation. Minneapolis: University of Minnesota. (Unpublished doctoral dissertation)

Owen, R.J. (1969). A Bayesian approach to tailored testing. (Research Bulletin RB-69-2). Princeton N.J.: Educational Testing Service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of American Statistical Association. 70, 351-356.

Patience, W.M., & Reckase, M.D. (1979). Operational characteristics of a one-parameter tailored testing procedure (Research report 79-2). Tailored testing research laboratory. Missouri University. (ED 198471)

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut.

Rosso, M.A., & Reckase, M.D. (1981) A comparison of a maximum likelihood and a Bayesian ability estimation procedure for tailored testing. Paper presented at the Annual Meeting of the National Council of Measurement in Education. (ED 187760)

Samejima, F. (1969). Estimation of latent abilty using a respone pattern of graded scores. Psychometrika. Monograph Supplement. 17

Weiss, D.J. (1980). Computerized adaptive performance evaluation: Final report, February 1976 through January 1980.. Minneapolis: Department of Psychology, University of Minnesota.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement. 6, 473-492.

Weiss, D.J., & Betz, N.E. (1973). Ability measurement: Conventional or adaptive. (Research Report 73-1). Minneapolis: Department of Psychology, University of Minnesota.

Weiss, D.J., & McBride, J.R. (1983). Bias and information of adaptive Bayesian testing. (Research report RR-83-2). Minneapolis: Department of Psychology, University of Minnesota.

Wood, R. (1972) Response contingent testing. Review of Educational Research. 43, 529-544.

Wright, B.D., & Masters, G.N. (1982) Rating Scale Analysis: Rasch Measurement. Chicago: MESA Press.