Two new Bayesian item selection methods for CAT:

Maximum Potential Impact (MPI) and Attraction Measure (AM)

Nadine Vestering[1,2]

Stef van Buuren[1]

Submitted, Sept 2004

[1] Department of Statistics, TNO Prevention and Health, P.O. Box 2215, 2301 CE LEIDEN

[2] Dept of Research Methodology, University of Leiden

Address for correspondence: S. van Buuren (S.vanBuuren@pg.tno.nl)

Two new Bayesian item selection methods for CAT:

Maximum Potential Impact (MPI) and Attraction Measure (AM)

Abstract

Two new Bayesian item selection methods for the use in CAT, Maximum Potential Impact (MPI) and Attraction Measure (AM), are introduced and compared with a popular current method, the Fisher information (FI) method. The MPI method can be used with any IRT model, the AM is specific to the Partial Credit Model. Extensive simulations showed that the three methods performed very similar. The FI method had a tendency to select polytomous items, and therefore usually required fewer items to reach certain reliability. The MPI and AM methods tended to favour dichotomous items, and the AM had smallest bias of the three methods, but at increased standard error. Also, the MPI and AM methods performed somewhat better for larger item banks, although requiring more items. Using a more informative item bank, like a polytomous item bank, had a positive influence on every aspect of performance of the MPI and AM methods. The AM method had a slightly advantage over both others when the test was stopped early, which makes the AM method attractive for applications where tests are very short.

1 Introduction

Item selection is an essential component in Computerized Adaptive Testing (CAT). Several item selection criteria and methods have been proposed over the years. Classic methods are generally restricted to item banks that contain dichotomous items only, and only recently have attempts been made to select items from item banks with polytomous or mixed dichotomous/polytomous items. Dodd, De Ayala & Koch (1995) and Berger & Veerkamp (1997) provided methods for the polytomous case using the Fisher information function. Van Rijn, Eggen, Hemker & Sanders (2002) studied the performance of such methods.

Van der Linden (1998) suggested Bayesian methods that might have superior properties. He introduced the notion of the preposterior distribution. In Bayesian analysis, the posterior is the distribution of the latent ability of an examinee *after* the examinee has given a response. The *pre*posterior is almost the same, except for the fact that it is calculated for each possible response alternative *before* the examinee has responded. Thus, each response alternative has its own preposterior. The comparison of preposteriors provides the scientific basis for choosing among different items. Though the ideas were presented for the dichotomous case, it is straightforward to extend them to polytomous items.

The present paper introduces two new item selection methods for polytomous data, both based on preposteriors. The first method is called Maximum Potential Impact (MPI). It is a general Bayesian method that does not depend on any particular psychometric model. The second method, called Attraction Measure (AM), relies on a specific property of the Partial

Credit Model (PCM; Masters, 1982). We evaluated the performance of both new methods in terms of bias and efficiency, and compared their properties to Fisher information (FI) method, arguably the most popular method for selecting items from polytomous item banks.

The paper first introduces new methods, and then describes a simulation design used to evaluate their performance. This is followed by a presentation the results. The discussion summarizes the main results and provides some additional points.

2 Item selection methods

2.1 Preliminaries: Calculation of the preposterior density

Let there be $q > 1$ candidate items ($j = 1,..., q$) in the item bank, each with $m_j > 1$ response categories ($k = 1,\ldots, m_j$). Let $?$ denote the continuous trait to be measured. Two ingredients are needed for calculating the preposterior distribution for each response alternative k. First, we need to know how the probability of responding in each category varies with $?$, i.e. $P(Yj=k|?)$. This function is typically specified by the parameters of the IRT model used. Second, we need the current distribution of ability of an examinee, $P(?)$. This is typically available from the previous examinee responses. If no items have yet been administered, $P(?)$ can for example be taken as a uniform distribution over the interval [-10,+10].

The preposterior of response category k in item j is the distribution of $?$ when the examinee would answer that category. This can be calculated by Bayes rule as

$$P(?|Y_j=k) = \frac{P(Y_j = k|q)P(q)}{\sum_j P(Y_j = k|q)P(q)}, \quad (1)$$

evaluated on a grid of *?*-values. Note that (1) is completely general and does not depend on a specific IRT model. There are m_j preposteriors $P(?|Y_j=k)$, one for each category, and these represent the potential posterior proficiency that results if the examinee would answer response category *k* on item *j*.

--- INSERT FIGURE 1 ABOUT HERE ---

Figure 1 illustrates the key concepts. Suppose the item bank contains a four-category item with thresholds -1.776, -0.014 and 4.493 fitted according to the PCM. Figure 1a present the Item Characteristic Curves $P(Y_j=k|?)$ of this item. Note that at each *?* the probabilities add up to 1. Suppose also that current ability distribution $P(?)$ of the examinee is given by Figure 1b. The area under distribution is standardized to 1, so $P(?)$ can be interpreted as a density. Figure 1c contains the corresponding preposterior distributions according to equation (1). Also here the category preposteriors $P(?|Y_j=k)$ can be interpreted as densities. The preposterior density $P(?|Y_j=k)$ indicates how the current ability P(?) changes if the respondent would respond category *k* of item *j*. A response in category 1 would sharpen the ability estimate and shift it slightly towards the left, whereas a response in category 4 would seriously alter both the location and the spread of the ability distribution. Observe that for this respondent a response in category 4 is very unlikely though.

2.2 Maximum Potential Impact (MPI)

For a given $P(?)$, we calculate the category preposteriors $P(?|Y_j=k)$ for all candidate items. Items whose category preposteriors are similar should not be selected because the ability distribution will hardly change as a result of the respondent's answer. We quantify the potential impact an item can have on the preposterior by the amount in which the m_j preposteriors differ, i.e., larger between-category variation results in a higher selection probability. The method selects the item that has Maximum Potential Impact (MPI).

Two complexities arise. First, note that the impact measure should account for the relative likelihood of each category as response probabilities will depend on $P(?)$. Second, when dealing with mixed dichotomous/polytomous items banks, there should be some way of comparing items with a different number of categories.

We address both difficulties analysis of variance on the m_j preposterior densities, with group sizes depending on $P(?)$. Let $\hat{q} = E[P(?)]$ be the Expected A Posteriori (EAP) estimate of $?$, and let $n_k = nP(Y_j=k|?=\hat{q})$ be a virtual group size proportional to the expected probability of response k, with $n > 1$ taken as an arbitrary scaling constant representing 'total sample size'. Furthermore, define $m_k = E[P(?|Y_j=k)]$ as the mean of the k'th preposterior, $s_k^2 = VAR(P(?|Y_j=k)$ as its variance, and $m = \sum n_k m_k / n$ as the grand preposterior mean. Then SS(between) = $\sum n_k (m_k - m)^2$ and SS(within) = $\sum (n_k - 1) s_k^2$, so MS(between) = SS(between)/(m_j - 1) and MS(within) = SS(within)/(n - m_j). A convenient measure of impact is then F = MS(between) / MS(within). Under the usual assumptions of the F-test, this

measure follows an F-distribution with (m_j - 1) and (n - m_j) degrees of freedom, so $p_j = F(m_j - 1, n - m_j)$ is the associated significance level. The MPI method selects the item with smallest p_j.

We take the scaling constant as n = 50 for convenience. As long as n is much larger than k, the exact choice of n will not affect the relative order of F-values (or of their associated p-values). Consequently, item selection is insensitive to the choice of n.

2.3 Attraction Measure (AM)

Under the Partial Credit Model and Bayes rule (1), the marginal probability of response category k under the current prior $P(?)$ is proportional to the ratio of $P(Y_j = k|?)$ and $P(?|\ Y_j = k)$, i.e.

$$P(Y_j = k) \propto \frac{P(Y_j = k|q)}{P(q|Y_j = k)} \qquad (2)$$

for any $?$. This is a surprising but convenient property that allows us to calculate the expected marginal frequency distribution $P(Y_j=k)$ of item j under $P(?)$. $P(Y_j=k)$ measures the attractiveness of each response category.

The idea behind the Attraction Measure (AM) method is that items with a more homogeneous $P(Y_j=k)$ are more informative, and hence should obtain a higher priority of being selected. For

example, dichotomous items with $P(Y=1) = P(Y=2) = 0.5$ are more informative at $P(?)$ than those with $P(Y=1)$? $P(Y=2)$. The ideal distribution $(1/m_j)$ is defined as the one in which each response category has an equal probability of being selected. A convenient information measure is the variance among $P(Y_j=k)$, i.e., $s_j^2 = (1/(m_j - 1))\sum (P(Y_j = k) - (1/m_j))^2$. The AM method selects the item the smallest s_j^2.

2.4 Fisher Information (FI)

The Fisher information (FI) method is probably the most popular item selection methods. Lord (1980) seems to be the earliest reference. Under the Generalized Partial Credit Model (Muraki, 1992), the Fisher information of item *j* can be calculated as (Donoghue, 1994)

$$I_j = a_j^2 \left[\sum_{k=1}^{m_j} k^2 P(Y_j = k \mid \hat{q}) - \left(\sum_{k=1}^{m_j} kP(Y_j = k \mid \hat{q}) \right)^2 \right]. \quad (5)$$

Setting a_j =1 yields the Partial Credit Model. The FI method selects the item with largest I_j.

3 Evaluation

3.1 Simulation design

A Monte Carlo simulation study was done to evaluate the performance of both new methods. Some conditions were common to all experiments. The EAP method (Bock & Mislevy, 1982) was used to estimate the latent ability. For dichotomous items, simulated item thresholds for dichotomous items were drawn from N(0, 2.5). Similarly, we use for three categories N(-1, 2)

and N(1, 2); for four categories N(-2, 1.5), N(0, 1.5) and N(2, 1.5); and for five categories N(-2.25, 1.25), N(-0.75, 1.25), N(0.75, 1.25) and N(2.25, 1.25). Respondents were generated at eleven ability levels -5 through +5 with a grid size of 1. At each level, 500 respondents were generated, so each experimental condition was tested by N=5500 respondents. All calculations were done by customized programs in S-Plus 6.2.

The following experimental factors were systematically varied:

- *Item selection method (M)*: Three item selection methods described in section 2 were implemented: MPI, AM and FI.
- *Size of the item banks (S)*: Three sizes of item bank were specified: 25, 50 and 100 items.
- *Number of categories (C)*: The number of categories per item was specified as 2, 4 and as a mixture of items with equal numbers of 2, 3, 4 and 5 categories. The latter is called the mixed bank.
- *Termination rule (R)*: The test was terminated when ability was estimated with certain reliability *?*. Three reliability levels were specified: 0.7 (low), 0.8 (medium) and 0.9 (high). In practice, SE is the standard deviation (SD) of posterior distribution , and the test was terminated with SE's smaller than 0.55, 0.45 and 0.32, which according to the relation $? = 1 - SE^2$ correspond to reliability levels of respectively 0.7, 0.8 and 0.9.

Outcome variables were bias, standard error (SE) and number of items administered. Bias refers to the difference between the true and estimated ability. It is a measure of the accuracy of the estimation. For a sample of size N, bias is defined by

$$\text{Bias}(\hat{q}) = \frac{1}{N}\sum_{r=1}^{N}|q_r - \hat{q}_r|, \quad (3)$$

where $\hat{q}_r$ is the EAP estimate for the r^{th} replication and N is the number of replications. The SE is the spread among ability estimates, is a measure of the stability of the estimation, and defined as

$$\text{SE}(\hat{q}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{q}_r - \frac{\sum_{t=1}^{N}\hat{q}_t}{N}\right)^2} \quad . \quad (4)$$

3.2 Analysis

The data from the experiment were analysed in several ways. For the purpose of the analysis, we derived a factor termed *Distance (D)* as the absolute distance to the midpoint (zero) of the scale. The levels of *D* consist of the consecutive integers between 0 and 5. In addition, we derived a factor termed *Test Length (T)* with 2 levels indicating that the CAT was terminated after a fixed number of 5 items (level 1) or 10 items (level 2).

Several different ANOVA's were carried out. We first concentrate on the differences between the three item selection methods at the beginning of the test. This is followed by similar analyses at the end of the test. Both comparative analyses are done only on the largest item

bank size (100). Next, the properties of the MPI and AM methods will be studied in more detail.

For simplicity, three- and higher-way interactions are left out of the analyses. The sample sizes in the simulation are chosen in advance, so the traditional significance statistics are somewhat artificial, but can still be useful in tracing the relative order of effects. The effect size of a factor will be expressed in terms of $?^2$, the proportion of variance explained by all categories of the factor. Estimated means will be reported for selected main and interaction effects.

4 Results

4.1 Comparison of three methods

--- INSERT TABLE 1 ABOUT HERE ---

Table 1 displays the outcomes for the three methods of interest when the test is terminated after 5 or 10 items. For a short test of 5 items, the AM is less biased than the other two methods, at the expense of a larger SE. The effect persists at 10 items, though it is somewhat attenuated.

--- INSERT TABLE 2 ABOUT HERE ---

Table 2 shows the same phenomenon for longer tests, but overall the differences between the three item selection methods are small. As in Table 1, Table 2 shows that the performance of the methods is quite similar for larger number of items. Note that the new MPI and AM methods require slightly more items than the FI method. An unexpected finding is that bias slightly increases as more items are administered. One possible explanation is that this occurs because the item bank becomes exhausted when it is used to assess people at the extremes of the scale.

--- INSERT TABLE 3 ABOUT HERE ---

Table 3 displays the results of the ANOVA using *M*, *C*, *T* and *D* as factors, under both fixed test length (5 or 10 items) and fixed test reliability (0.7, 0.8, 0.9) for the three outcome measures. The effect size $?^2$ for the *M* factor (0.016) and its interactions are small in comparison to the effect sizes of the other factors, so differences between item selection methods are comparatively small. There is an *M*×*C* interaction. Closer inspection of the data reveals that the three methods do not differ when a dichotomous item bank is used. This is reasonable because all item selection methods order the dichotomous items according to the distance between the prior mean and the item parameter. For polytomous and mixed item banks, the methods do not differ in SE's, but as before, AM shows less bias after administering 5 items.

The performance of MPI is generally in between that of AM and FI. For a stopping rule based on test reliability, the methods hardly differ in bias and SE, only in the number of items used. The FI method needs fewer items when a mixed item bank is used. In general, the AM shows

less bias at the extremes of the scale when 5 or 10 items are administered, at the expense of larger SE's in the middle of the scale. For fixed test reliability, no differences related to distance from the mid point appear, except for the number of items used. The MPI and AM require more items than the FI method on all distances studied.

--- INSERT TABLE 4 ABOUT HERE ---

Table 4 shows a result with important practical implications. First observe that the FI method uses the smaller total number of items (25.4). Also, note that the Fisher exhibits a preference for items with more categories, whereas by contrast the MPI clearly favours dichotomous items. In general, the strategy employed by the FI method is the better one when the costs of answering a two-category item and a five-category item are equal. In that case, using FI is the more efficient. If however, the administration of dichotomous items is more efficient than polytomous items, the MPI method becomes more interesting. It may require more items, but the total administration time could still be shorter than the test selected by the FI. Furthermore, we found that the AM method generally has less bias than the other two methods, but has higher SE's. The AM method may therefore be of interest in applications where small bias is more important than small SE's.

4.2 MPI: Detailed analysis

--- INSERT TABLE 5 ABOUT HERE ---

The upper part of Table 5 shows the results of ANOVA with the MPI method. Distance to the midpoint of the scale is by far the most important factor for all outcomes accounting for 40.7 %, 33.6 % and 25 % of the total variance respectively. The means of the different levels of *D* vary considerably, with the most favourable outcomes associated with small values of *D*. Thus, the method works best in the middle of the scale, which is consistent with other research (e.g., van der Linden, 1998; Wang & Wang, 2001; van Rijn *et al*, 2002).

--- INSERT FIGURE 2 ABOUT HERE ---

Figure 2 is helpful device in interpreting the ANOVA results. It displays the estimated means of each level of all main factors on the outcomes. An interesting result is that a higher specified reliability is not associated with bias. The relation between reliability and SE and number of items is as expected. Thus, a longer test will make the estimate more stable, but does do little to improve accuracy. In fact, Figure 2 suggests that using a longer test might even be harmful for accuracy.

Donoghue (1994) showed that items with more categories are more informative in the sense that they reduce the standard error of the ability estimate. A rough measure of the informativeness of an item bank is the total number of the categories of the items. Therefore, we expect that an item bank consisting of polytomous items will have a better performance than an item bank with the same number of dichotomous items only. Figure 2 clearly illustrates this effect. The item bank with items of four categories ($4 \times 100 = 400$) is most informative and performs best. The mixed item bank is also very informative ($2 \times 25 + 3 \times 25 + 4 \times 25 + 5 \times 25 = 350$) and the dichotomous item bank is least informative ($2 \times 100 = 200$).

The differences between the mixed item bank and the item bank with items of four categories are small and only result in the mixed item bank requiring more items.

Also, the size of the item bank (i.e., 25, 50 and 100) has an effect on the performance of the MPI method. The method performs better with larger item banks in terms of bias and SE, but requires more items than with a smaller item bank.

Table 5 also contains interaction effects. Since not all of the effects are of immediate importance to the performance of MPI, the discussion of results will be limited to just a few interactions.

The interaction between size of the item bank and distance (S × D) has the largest effect on the outcomes. This effect was not expected a priori. At the midpoint of the scale the performance of the item banks is similar. Although the larger item banks have more choice in items they perform equally well. Apparently having more choice only pays off primarily for ability levels further away from the midpoint. At three or more theta units away from the midpoint, the method using the larger item bank shows less bias, has smaller SE's while using more items.

Another substantial effect concerns the interaction between size of the item bank and number of categories (C × S). If the MPI method uses a smaller item bank, the dichotomous and mixed item bank show more bias, larger SE's and require fewer items. More bias and larger SE's is also shown for the item bank containing items with four categories when a smaller

item bank is used. However the number of items required is independent of size of the item bank.

Apart from some unexpected effects discussed above, the MPI performs overall as expected in the specific situations.

4.3 AM: Detailed analysis

--- INSERT FIGURE 3 ABOUT HERE ---

The lower half of Table 5 displays the ANOVA results for the AM method. In general, the results conform to those for the MPI method. Comparing Figure 2 and 3, it is clear that also the means of the levels of the factors are very similar. This also extends to the interaction effects, and indicates that the factors influence the performance of the MPI and AM in the same way. Thus, the results of the MPI method apply to the AM method as well.

5 Discussion

We proposed two new Bayesian item selection methods, both based on the preposterior. The new methods were compared with the method using Fisher information. Overall, all methods perform quite similar, but there are some small differences. The FI has a tendency to select polytomous items and therefore usually requires fewer items to reach certain reliability. The MPI and AM tend to favour dichotomous items, and the AM has smallest bias of the three

methods, but at increased standard error. Also, the MPI and AM perform somewhat better for larger item banks, although requiring more items. Using a more informative item bank, like a polytomous item bank, had a positive influence on every aspect of performance of the MPI and AM. This corresponds to other research where item selection methods were tested with different item banks.

As noted, there is a trade off between the accuracy and the stability of the estimate. The AM method is less biased than the MPI and FI at the beginning of the test, but with larger standard errors. This trade off appears to be related to the number of categories of the selected items. The AM and MPI methods prefer dichotomous items. Apparently, using dichotomous items restricts the estimate less than polytomous items, i.e., dichotomous items only pull the estimate towards one side of the scale. In contrast, polytomous items work more on a range of the scale and the estimate increases more in stability than with dichotomous items. Thus, if a quick assessment of the extremes of the scale is needed, using dichotomous items employing the AM may be preferred. On the other hand, a reasonable across-the-board estimate could better be made by administrating polytomous items with the FI method.

Although we have tried to design a realistic experiment, things might turn out differently in practice. The estimation of the item parameters based on real data is never free of error. According to van der Linden & Glas (2000) the errors in item parameters can have a negative effect on CAT estimation. The present study assumed that the item parameters were true. We feel though that the simulation study gives a good indication of the relative performance of item selection methods. In a real situation, an examinee can give deviating answers at the

beginning of the test, because the examinee may need to adjust to taking a CAT. Such effects could not be incorporated in the study entirely, but need careful attention in practice.

It would be interesting to compare the performance of the methods using real data. We think it is important to examine the performance of the MPI and AM methods more thoroughly, because both may have practical advantages over the FI method. The MPI method can be used for any IRT model, and thereby eases the use of item banks made with other IRT models. On the other hand, the MPI needs more computational time than either the FI or the AM methods. Another point of interest is the hybrid methods. Perhaps, results could improve by starting with a couple of dichotomous items to get a rough assessment, before using polytomous items.

The simulation study is quite broad. It not only addresses three different item selection methods, but also deals with several factors that influence the performance of CAT. We found that the two new Bayesian item selection methods are performing as well as the FI method. We look forward to apply these methods in practice.

References

Berger, M.P.F. & Veerkamp, W.J.J. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, 22, 203-226.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.

Dodd, B.G., De Ayala, R.J., & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19, 5-22.

Donoghue, J.R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the Generalized Partial Credit Model. *Journal of Educational Measurement*, 31, 295-311.

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

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, 25, 317-331.

Van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63, 201-216.

Van der Linden, W.J. & Glas, C.A.W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, 13, 35-53.

Van Rijn, P.W., Eggen, T.J.H.M., Hemker, B.T., & Sanders, P.F. (2002). Evaluation of selection procedures for Computerized Adaptive Testing with polytomous items. *Applied Psychological Measurement*, 26, 393-411.

Table 1: Mean bias and SE of three item selection methods after a fixed test length of 5 and 10 items for an item bank size of 100.

Fixed Test Length	Method	Bias	SE
5 items	MPI	0.132	1.00
	AM	0.098	1.07
	FI	0.148	0.94
10 items	MPI	0.080	0.64
	AM	0.073	0.68
	FI	0.079	0.62

Table 2: Mean bias, SE and number of items needed of three item selection methods under various stopping rules for an item bank size of 100.

Fixed Test Reliability	Method	Bias	SE	Number of items administered
0.7	MPI	0.030	0.53	16.7
	AM	0.027	0.54	16.7
	FI	0.029	0.53	15.3
0.8	MPI	0.034	0.46	27.9
	AM	0.027	0.46	27.9
	FI	0.034	0.46	26.2
0.9	MPI	0.041	0.37	55.2
	AM	0.040	0.36	54.1
	FI	0.040	0.36	52.4

Table 3: Proportion explained variance per effect from ANOVA for early (fixed length) and late termination rules (fixed reliability). The item bank size is 100.

Stopping rule	Effect	DF	bias	SE	# items
Fixed length	- Main Effects				
	Methods (M)	2	.010	.016	-
	Categories (C)	2	.077	.292	-
	Test Length (T)	1	.047	.369	-
	Distance (D)	5	.303	.164	-
	- Interactions				
	M × C	4	.008	.037	-
	M × T	2	.006	.003	-
	M × D	10	.018	.004	-
	Error	171			
Fixed reliability	- Main Effects				
	Methods (M)	2	.001	.000	.001
	Categories (C)	2	.088	.086	.181
	Reliability (R)	2	.008	.044	.280
	Distance (D)	5	.327	.218	.373
	- Interactions				
	M × C	4	.001	.002	.001
	M × R	4	.000	.000	.000
	M × D	10	.002	.002	.000
	Error	267			

= statistically significant

Table 4. Selection of items from a mixed item bank. Given are the total number of items administered (bottom row) and the breakdown into the number for 2, 3, 4 and 5-category items.

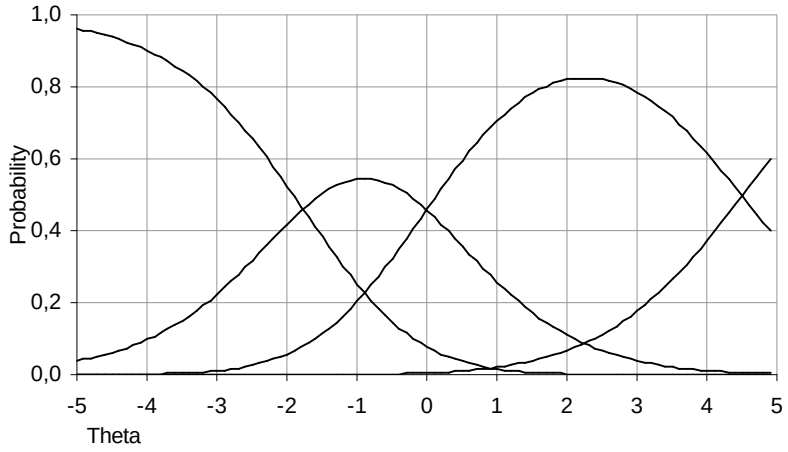
Categories	MPI	AM	FI
2	9.9	8.8	3.6
3	8.5	4.9	4.9
4	6.6	6.5	6.5
5	5.9	9.3	10.4
Total	30.9	29.5	25.4

Table 5. Proportion explained variance per effect from ANOVA for the MPI and AM item selection methods.

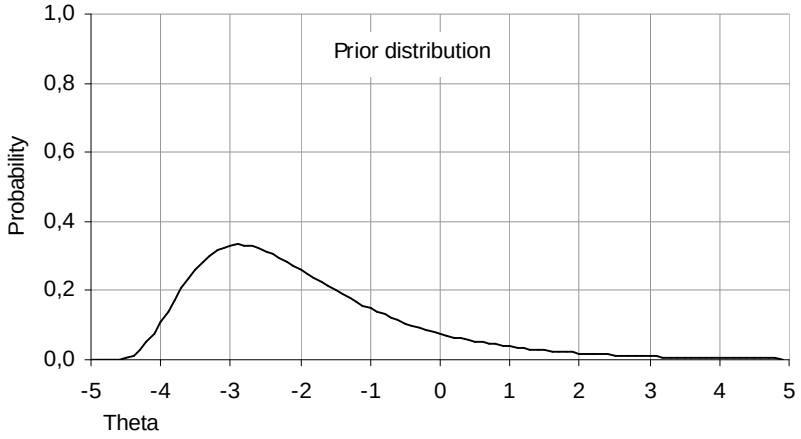
Method	Effect	DF	bias	SE	# items
MPI	- Main Effects				
	Categories (C)	2	.067	.127	.128
	Item bank Size (S)	2	.083	.139	.060
	Reliability (R)	2	.000	.046	.177
	Distance (D)	5	.407	.336	.250
	- Interactions				
	C × S	4	.022	.038	.034
	C × R	4	.000	.004	.002
	C × D	10	.096	.050	.004
	S × R	4	.000	.006	.069
	S × D	10	.149	.090	.100
	R × D	10	.002	.006	.009
	Error	243			
AM	- Main Effects				
	Categories (C)	2	.064	.122	.135
	Item bank size (S)	2	.098	.140	.060
	Reliability (R)	2	.000	.049	.172
	Distance (D)	5	.383	.335	.240
	- Interactions				
	C × S	4	.028	.047	.036
	C × R	4	.000	.004	.002
	C × D	10	.082	.042	.004
	S × R	4	.000	.007	.064
	S × D	10	.140	.100	.093
	R × D	10	.003	.006	.010
	Error	243			

= statistically significant

(a)


1,0
0,8
0,6
0,4
0,2
0,0
Probability
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta

(b)


Prior distribution
1,0
0,8
0,6
0,4
0,2
0,0
Probability
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta

(c)

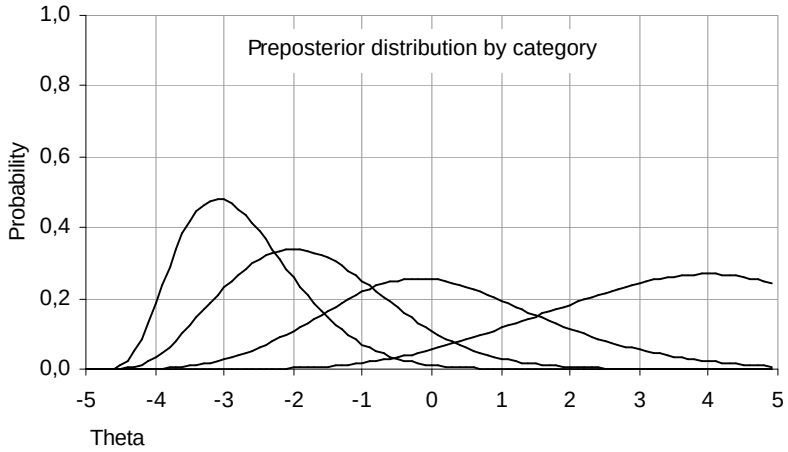

Preposterior distribution by category
1,0
0,8
0,6
0,4
0,2
0,0
Probability
-5 -4 -3 -2 -1 0 1 2 3 4 5
Theta

Figure 1

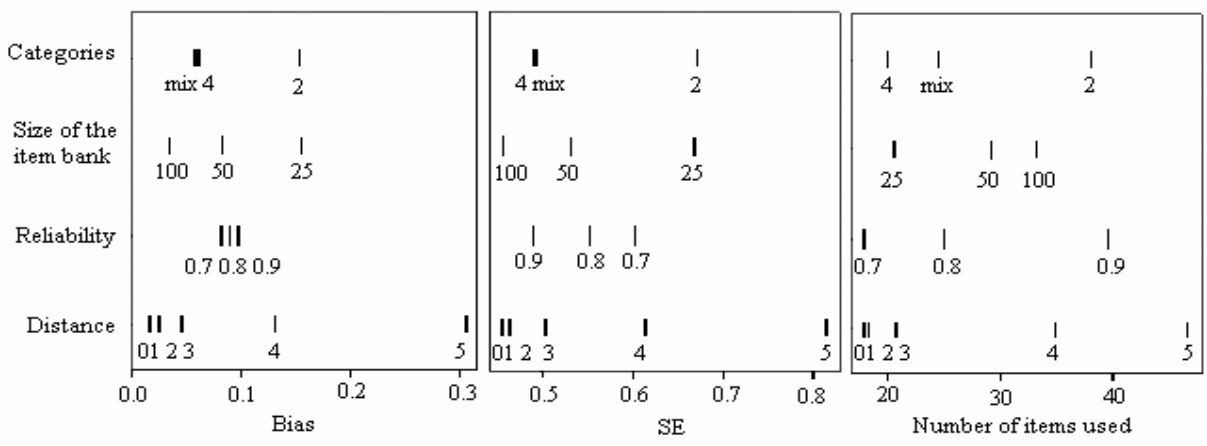

Categories
mix 4
2
4 mix
2
4
mix
2
Size of the item bank
100
50
25
100
50
25
25
50
100
Reliability
0.7 0.8 0.9
0.9
0.8
0.7
0.7
0.8
0.9
Distance
01 2 3
4
5
01 2 3
4
5
01 2 3
4
5
0.0
0.1
0.2
0.3
Bias
0.5
0.6
0.7
0.8
SE
20
30
40
Number of items used

Figure 2

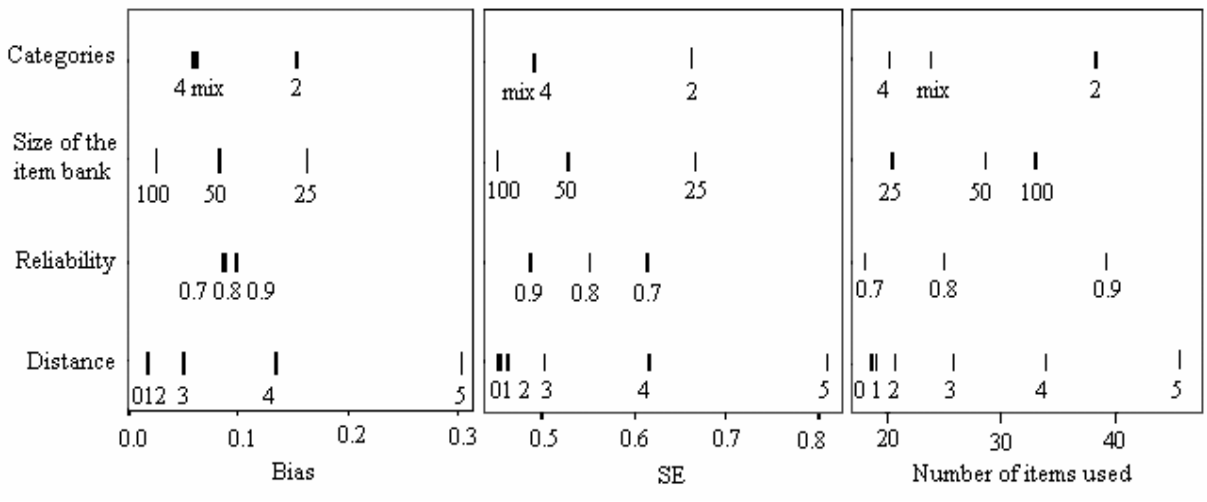

Categories
4 mix 2
mix 4 2
4 mix 2
Size of the item bank
100 50 25
100 50 25
25 50 100
Reliability
0.7 0.8 0.9
0.9 0.8 0.7
0.7 0.8 0.9
Distance
012 3 4 5
01 2 3 4 5
0 1 2 3 4 5
0.0 0.1 0.2 0.3
Bias
0.5 0.6 0.7 0.8
SE
20 30 40
Number of items used

Figure 3

Figure captions

Figure 1: Illustration of the preposterior distributions for the categories of a polytomous item.

Figure 2. Means of every level of the main factors on the outcomes bias, SE and number of items administered for the ANOVA for the MPI method.

Figure 3. Means of every level of the main factors on the outcomes bias, SE and number of items administered for the ANOVA for the AM method.