Measurement and Research Department Reports 91-7

Decision Accuracy In IRT Models

H.H.F.M. Verstralen
N.D. Verhelst

Cito

Measurement and Research Department Reports 91-7

Decision Accuracy In IRT Models

H.H.F.M. Verstralen
N.D. Verhelst

Cito
Arnhem, 1991

Abstract

A disadvantage of the application of the Rasch Model (RM) or the Partial Credit Model (PCM) is that one may be forced to omit the best discriminating items from an item bank to attain acceptable model fit. A solution to this problem is offered by the family of One Parameter Logistic Models (OPLM). OPLM opens the possibility to model differently discriminating items on one latent scale, without sacrificing sufficient statistics and conditional maximum likelihood estimation. The preservation of these valuable properties of the RM and the PCM is achieved by avoiding to estimate discrimination indices by treating them as known integer constants. Although a dedicated least squares algorithm helps the user to quickly find appropriate values for the discrimination indices, he is in principle burdened with the responsibility for them.

This paper explores how much the user gains by adopting an OPLM approach in terms of the probability of misclassifying a student with a certain latent parameter value. Although the exploration is restricted to binary scored items, several perspectives are considered. These are scoring (weighted vs. unweighted), test construction (use of discrimination indices or not) and calibration (use of OPLM or the RM). Moreover, to elucidate the practical relevancy of the investigation, a real item bank at the elementary school level is used. It turns out that when OPLM is fully used for calibration and test construction the application of raw scores in stead of weighted scores causes only a minor loss of accuracy. However, appreciable losses in accuracy are incurred if the test is constructed with disregard of the discrimination indices and even more so when the RM is used as the measurement model.

Key words: Logistic IRT models, Scoring, Test construction, Measurement accuracy.

Introduction

Until the development of the One Parameter Logistic Model (OPLM, see below), the psychometric practitioner had to choose between the Rasch Model (RM), or the Partial Credit Model (PCM) for multicategory items, and the Birnbaum Model (BM) for IRT item analysis. With the RM (and the PCM) he ran the risk of having to abandon the best discriminating items, and the BM is contaminated with severe problems in parameter estimation. By treating discrimination indices not as parameters to be estimated, but as known integer constants, the application of OPLM allows him to keep items of various discriminating power in the same calibrated item pool without running into estimation problems. The idea to consider discrimination indices as known constants is new in psychometric models and sounds as an unbearably harsh burden on the shoulders of the practitioner. Although new in psychometrics, in other fields of social science modelling like factor analysis it is common for the user to specify the dimensionality of the solution and in LISREL the user typically has to provide a host of restrictions and other model specifications. However, like in factor analysis with its elbow heuristic for the dimensionality, the doctrine of known constants is not applied in its pristine theoretical severity. The practitioner is offered the assistance of a dedicated least squares algorithm to provide him with reasonably good values for the discrimination indices. The estimation in OPLM can, therefore, be viewed as a two stage procedure. The first stage is a 'quick and dirty' method without statistical testing, where all unknown parameters, including the discrimination indices and person parameters, are simultaneously fitted. The second stage treats the discrimination indices from the first stage, rounded to the nearest integer, as known constants and estimates the item parameters by CML or MML. Elaborate practice at Cito has shown that most times item test statistics indicate that just a few discrimination indices need to be adapted by one point in a direction as shown by these same statistics.

Nevertheless, the practitioner has to bear the responsibility for the discrimination indices, and, therefore, it makes sense to investigate what

psychometric benefits are his reward in addition to saving the best discriminating items.

OPLM as a Formal Model

In OPLM the probability for student k of gaining a score j on item i with m_i as its maximum score and a_i as its discrimination index (a positive integer) is given by the expression:

$$P(x_{ki}=j \mid \theta_k, \underline{\eta}_i) = \frac{\exp(a_i(j\theta_k - \eta_{ij}))}{1 + \sum_{d=1}^{m_i} \exp(a_i(d\theta_k - \eta_{id}))}, \quad a_i \in \{1,2,....\}. \qquad (1)$$

If all discrimination indices a_i are equal then (1) reduces to the PCM, and if, moreover, all m_i are set to one, the RM is obtained. Therefore, it is not difficult to infer from (1) that OPLM is a generalisation of the RM and the PCM as well. A comprehensive introduction to the model, the statistical tests, and applications can be found in Verhelst a.o. (1991).

The investigation reported in this article is focused on the gain in measurement accuracy by having the discrimination indices vary over the small positive integers in stead of restricting them to be equal as in the Rasch Model. Therefore, to avoid distracting complexity, the current application of OPLM is restricted to binary items. Consequently, (1) may here be simplified to:

$$P(x_{ki}=1 \mid \theta_k, \eta_i) = \frac{\exp(a_i(\theta_k - \eta_i))}{1 + \exp(a_i(\theta_k - \eta_i))}, \quad a_i \in \{1,2,....\}, \qquad (2)$$

the RM enriched with discrimination indices.

It follows from formulas (1) and (2) that the set of discrimination indices (a_i, i=1,2...) must also be interpreted as a scaling factor. Multiplication of the discrimination indices by an arbitrary factor c can be compensated for by dividing person and item parameters by the same factor c. This indeterminacy introduces a problem in interpreting the latent parameters.

Therefore, to support a consistent interpretation, we will sometimes refer to a 'standard' scale with a unit called 'standit'. This is defined to be the scale where the geometric mean (we deal with a multiplicative factor, not an additive one) of the discrimination indices equals 1, like in the RM. The standard scale is obtained by multiplication of the latent parameters by the geometric mean of the discrimination indices, and by division of the discrimination indices by this same number. In the standard scale the discrimination indices are, of course, not restricted to be integers.

Method of Investigation

The key concept of this investigation is a misclassification function. For a certain measurement condition to decide whether the ability of a student does exceed a prespecified level, the misclassification function gives the probability of the wrong decision. Six conditions will be compared which are combinations of the following aspects:

1. Calibration of the item bank with OPLM vs. RM
2. Decision based on weighted vs. raw score
3. For OPLM: test construction with or without regard of discrimination indices
4. For RM: Calibration of the entire item bank or of a more Rasch homogeneous subset.

Aspect 3 needs some clarification: An optimum test of length k for deciding whether the latent parameter of a student exceeds a certain level θ_0 contains the first k items from the ordered item bank according to information at θ_0, the item with highest information first (Verstralen & Verhelst, 1991). If the discrimination indices are neglected in calculating the information, the optimum test contains the k items with difficulty parameter closest to θ_0. With regard to this third aspect one might wonder for whatever reason one would choose to construct a test with neglect of discrimination indices, after taking the trouble of calibration under OPLM. The reason is that optimal test construction, in general, tends to use preferably high discriminating items, leaving the greater part of an item bank unused. One way of overcoming this is to proceed as described in aspect 3, and to construct tests using psychometric

information exclusively on the difficulty of items. By preserving the optimal test construction only for really important educational decisions, one safeguards the value of the better discriminating items from erosion by unnecessary frequent use.

These four aspects generate the scheme of test conditions as given in Table 1.

TABLE 1

Six test conditions to compare measurement accuracy

Condition	Acronym	Item bank Calibration	Discr. Indices in Test Construction	Score
1	ODW	OPLM	Yes	Weighted
2	ODR	,,	,,	Raw
3	ONW	,,	No	Weighted
4	ONR	,,	,,	Raw
5	RNR	RM	,,	,,
6	HNR	Homogeneous RM	,,	,,

When the RM is applied there obviously is no choice in using discrimination indices in test construction or scoring.

Because we value to give an impression of measurement accuracy under practical measurement conditions a real 'item bank' is used. The complete item bank contains 245 items on Reading Comprehension at the elementary school level (Staphorsius, a.o., 1991). An artificial item bank or a series of artificial item banks would have been, perhaps, more apt to highlight relations between, for instance, variability in discrimination indices and measurement accuracy. However, because the connection with real measurement conditions would have been obscure, the relevancy of the investigation would have been difficult to assess.

For the calculation of the misclassification functions OPLM is considered the model that truly reflects reality. Therefore, whatever the test condition the misclassification functions are calculated with parameter values estimated within OPLM. This special treatment of OPLM will easily elicit the criticism that this choice, by itself, will give OPLM the upperhand concerning measurement accuracy. However, in contradistinction to the Rasch model, also in

the more homogeneous condition 6 (HNR), OPLM fits the data quite well: Although a lot of observations were involved (12,000 students) OPLM could not be rejected on a statistical basis (5% level). Moreover, the estimation errors of the parameters are small: 0.006 through 0.025. Their size on the standard scale is obtained by multiplication with 4.5: 0.027 through 0.113. These estimation errors are so small that substitution of the parameter estimates for their true values, which are not known, could not be expected to cause a notable difference in the misclassification functions.

After the first fitting stage the discrimination indices are taken to range from 2 through 9, with the distribution given in Table 2.

TABLE 2

Distribution of discrimination indices in the item bank

Discr. Index	2	3	4	5	6	7	8	9
Frequency	6	27	82	74	40	12	3	1

Geometric Mean 4.55 = 1/0.22

For test condition 6 (HNR), the more homogeneous Rasch bank, only the items with discrimination indices 4,5 and 6 are retained, 196 in total. The complete item bank, and this subset as well, are calibrated separately under the Rasch model.

To initially avoid the equating problem between OPLM and the RM it is easiest to start with the psychometrically optimal test condition 1 (ODW).

First the complete item bank with 245 items is calibrated with OPLM. Three decision levels θ_{0i} (i = 1,2,3) are chosen: at -1, 0, and 1 standit. Because the geometric mean of the discrimination indices in this bank is 1/0.22, these values have to be transformed to -0.22, 0, and 0.22. For each of them the same procedure is followed. Take, for instance, θ_{01} = -0.22. Because the 'item bank' we employ is relatively small compared to the size a real item bank is supposed to have, a test of only 16 items is constructed with optimal decision accuracy at this value of the latent trait. This amounts to selecting

the 16 most informative items at θ_{01}. To get tests of more customary length these 16 items are taken two or four times to generate tests of 32 and 64 items resp. For such a test the expected weighted score w_{01} at θ_{01} is calculated. If a student with latent parameter θ earns a score equal to or larger than the criterion score w_{01} he passes the exam else he fails. Whether this decision is correct depends on θ being larger resp. smaller than θ_{01}.

Now consider a test from the item bank with k items. Let $\underline{x}$ denote a response pattern with k elements: $x_i = 1$ if the response to item i is correct else $x_i = 0$. To compute the probability of an incorrect decision based on the weighted score w its distribution given θ is calculated by summation over all response patterns $\underline{x}$ that give rise to the same weighted score w, as shown in (3):

$$P(w|\theta) = \sum_{\{\underline{x}:(\underline{x},\underline{a})=w\}} \prod_i p_i^{x_i}(1-p_i)^{1-x_i} . \tag{3}$$

Here $(\underline{x},\underline{a}) = \sum_{i=1}^{k} x_i a_i$ denotes the inner product of the vectors $\underline{x}$ and $\underline{a}$, and p_i equals the probability of a correct response to item i given θ as expressed by (2). Summation in (3) is done over exactly those response patterns $\underline{x} \in \{0,1\}^k$ for which the inner product with $\underline{a}$ equals w.

To indicate the way in which these probabilities are calculated, we introduce the following notation: Let $\underline{\delta} \in R^k$ be a vector of reals, and $\underline{d} \in N^k$ a vector of nonnegative integer weights, both of dimension k. Then define:

$$\gamma_w(\underline{\delta},\underline{d}) = \sum_{\{\underline{x}:(\underline{x},\underline{d})=w\}} \prod_i \delta_i^{x_i} . \tag{4}$$

$\gamma_w(\underline{\delta},\underline{d})$ is called the basic combinatorial function of order w. Now, let $\delta_i = \exp(a_i(\theta - \eta_i))$, then we have from (3) and (4):

$$P(w|\theta) = \frac{\gamma_w(\underline{\delta},\underline{a})}{\sum_v \gamma_v(\underline{\delta},\underline{a})} , \tag{5}$$

where $\underline{a}$ denotes the vector with discrimination indices.

The probability of misclassification for $\theta < \theta_{01}$ is given by the sum of the probabilities (3) for all scores $w > \text{round}(w_{01} + 1)$ plus an interpolated part of the probability for $w = \text{round}(w_{01})$. Round(.) denotes the function that yields the nearest integer of its argument. The additional interpolated part equals $\text{round}(w_{01}) + 0.5 - w_{01}$. It results from a so called 'correction for continuity' when the 'continuous' scores are supposed to be uniformly distributed in the unit interval [w-0.5,w+0.5] around their observed integer value w. If all scores are considered continuous and uniformly distributed in the unit intervals around their observed nearest integer values, with total probability for the unit interval as given by (3), the above argument can be stated in a simpler way: The probability of misclassification given that $\theta < \theta_{01}$ equals the integral of the conditional density given θ of the scores $w > w_{01}$. Denote the value of this integral by SP, then, if $\theta \geq \theta_{01}$ the probability of misclassification is, of course, given by 1 - SP (see Figure 1).

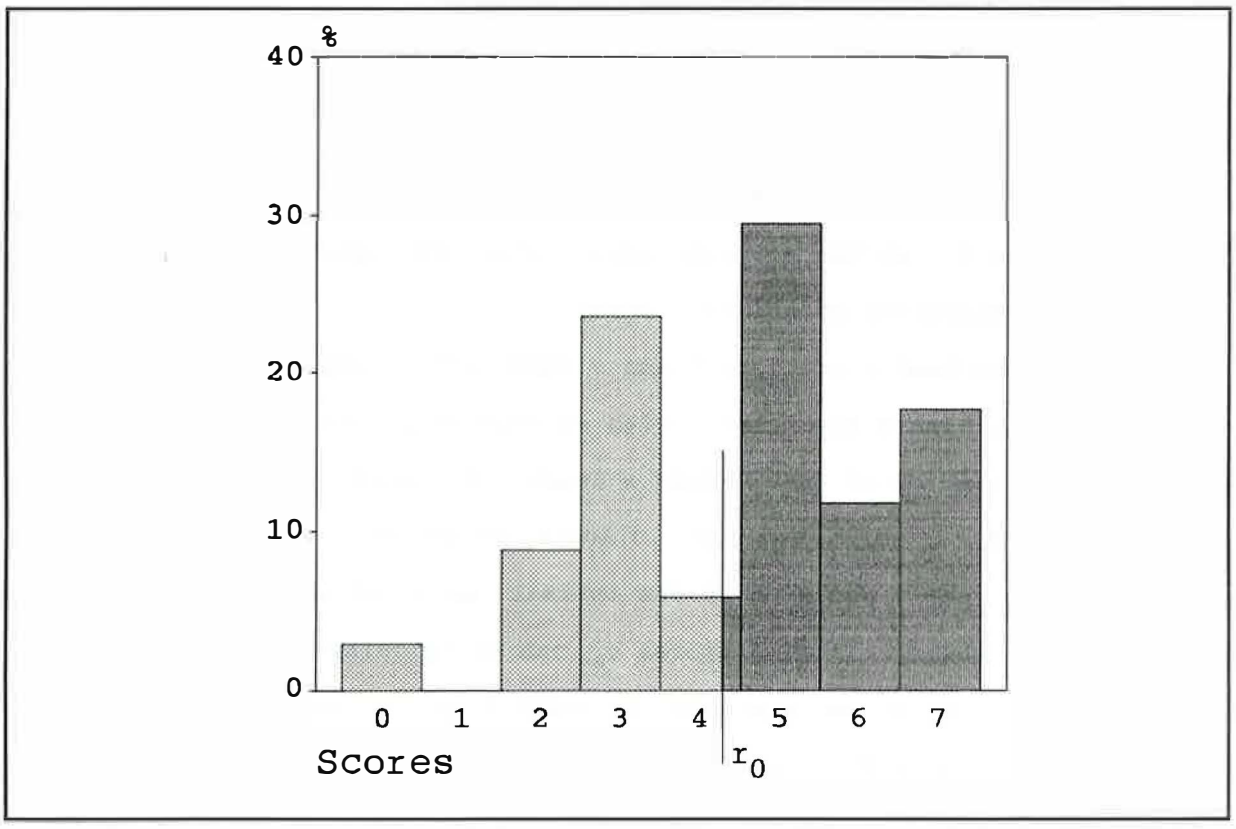

Figure 1. The darker area represents the probability of misclassification for $\theta < \theta_{01}$.

Calculation of the probabilities of misclassification for a closely spaced series of values for θ in a relevant range around θ_{01} gives an impression of the misclassification function. We chose 41 values in the interval $[\theta_{01} - 0.25, \theta_{01} + 0.25]$, which, for this item bank, is equivalent to about 2.25 standits. To be able to display possible discontinuities at θ_{01}, this value is represented by two values, one slightly smaller and one slightly larger than θ_{01}.

The second test condition (ODR) differs from the first (ODW) in that the unweighted raw score r is used instead of the weighted score w to decide pass or fail. The criterion score r_{01} is the expected raw score at θ_{01}. The conditional distribution of raw scores given θ is calculated very much like (5), the difference being that this time summation is done over all response patterns that give rise to the same raw score r:

$$P(r|\theta) = \frac{\gamma_r(\underline{\delta}, \underline{1})}{\sum_v \gamma_v(\underline{\delta}, \underline{1})}, \qquad (6)$$

where $\underline{1}$ denotes the k-dimensional vector of 1's.

Test conditions 3 and 4 (ONW, ONR) only differ from the first two by tests with less information at the decision level, which is expected to result in less accurate measurement around that level.

For the Rasch test conditions 5 and 6 (RNR, HNR), however, there is the problem of how to link a Rasch scale with an OPLM scale. We could start with the practical situation of just having a Rasch calibrated item bank and decision levels ξ_{01} at -1.0, 0.0 and 1.0 logit, or standit, which is equivalent in this case. For these decision levels optimal tests can be constructed in the Rasch Model, and their criterion scores r_{01} can be calculated. Given the items and a value θ^* of the latent variable the conditional raw score distribution can be calculated in OPLM. Consequently the two probabilities of scoring lower or higher than the critical score are known. However, because there is as yet no decision level in OPLM, it is not known to which of the two classes θ^* belongs, and, therefore, which of the two probabilities represents the probability of misclassification. This problem can be resolved by taking as

decision levels latent values θ_{0i} in OPLM with the same average probability correct on the common items as ξ_{0i} at the Rasch scales:

$$\sum_j P_j(\xi_{0i}) = \sum_j P_j(\theta_{0i}), \quad (i=1,2,3), \tag{7}$$

where j ranges over the common items in the OPLM bank and the particular Rasch bank. Needless to say that $P(\xi)$ is calculated within the Rasch framework and $P(\theta)$ within OPLM.

However, for the simultaneous presentation of the OPLM and Rasch misclassification functions it is preferable to have the same OPLM decision levels for all conditions. Therefore, equation (7) will be solved in the other direction: θ_{0i} in OPLM are considered known (-0.22, 0.0 and 0.22), and (7) will be solved for ξ_{0i} in the Rasch model. This will result in decision levels for the Rasch conditions that slightly deviate from -1.0, 0.0 and -1.0, as shown below.

Results

Before presenting results on decision accuracy, first some data on the item banks and on the constructed 16 item tests. For the Rasch Model conditions 5 and 6 (RNR, HNR) the just mentioned ξ_{0i} are:

TABLE 3

Decision levels for the two Rasch Model conditions RNR and HNR

Level	1	2	3
Condition			
5 RNR	-1.0223	0.0134	1.0384
6 HNR	-1.0607	-0.0028	1.0539

As was to be expected the decision levels are about -1, 0, and 1 resp.

The discrimination indices, difficulty parameters, and critical scores of the 16 item tests of the four OPLM conditions are shown in Table 4.

TABLE 4

Data on the tests with 16 items under OPLM conditions

	O P L M		Standard Scale			
Cond	Disc	Diff	Disc	Diff	w_0	r_0
ODW/R						
1/2-	6.94(.90)	-0.15(.06)	1.53(.20)	-0.67(.27)	40.98(111)	6.03
0	7.06(.75)	-0.04(.08)	1.55(.16)	-0.19(.35)	64.93(113)	9.13
+	6.25(.43)	0.17(.07)	1.38(.10)	0.79(.30)	57.69(100)	9.14
ONW/R						
3/4-	4.88(1.76)	-0.22(.02)	1.07(.39)	-0.98(.08)	37.85(78)	7.85
0	4.75(1.52)	-0.00(.01)	1.05(.34)	-0.01(.06)	38.64(76)	8.09
+	4.94(0.83)	0.21(.03)	1.09(.18)	0.97(.12)	40.03(79)	8.11

The first column in Table 4 gives the test conditions for which the test applies. For instance, 1/2- indicates test conditions 1 and 2 (ODW/R construction in OPLM with information maximisation at the decision level) at the negative decision level ($\theta_0 = -0.22$). The next column contains the means and, between () the standard deviations of the discrimination indices used in the OPLM calibration. Then, the mean and standard deviation of the item parameters as estimated by OPLM. The next two columns give the same information but rescaled to the standard scale by correcting for the geometric mean (1/0.22) of the discrimination indices in the bank. Under w_0 the critical weighted scores are displayed, and between () the maximum weighted scores. The Column labelled r_0 contains the critical raw scores. The maximum raw score equals 16 in every case and is, therefore, omitted. It appears from Table 4 that for conditions 1 and 2 (ODW, ODR) the difficulty of the tests at θ_0 deviates from 50% correct. Especially at $\theta_0 = -0.22$ the test is rather difficult, while at the other two levels the tests are relatively easy.

Table 5 shows the same data as given in Table 4,but now for the 16 item tests constructed under the two Rasch model conditions (RNR, HNR), except for some changes to the last columns. Under 'Rasch Diff' the item parameters are given as estimated under the Rasch model for the relevant item banks. Moreover,

in addition to the critical scores as calculated in the Rasch model in column r_0, the column labelled $E(r|\theta_0)$ shows the expected score at the decision level for the assumed 'reality' in OPLM. Note that r_0 equals $E(r|\xi_0)$ in the Rasch model.

TABLE 5

Data on the tests with 16 items under Rasch conditions

	O P L M		Standard Scale		R A S C H		
Cond	Disc	Diff	Disc	Diff	Diff	r_0	$E(r\|\theta_0)$
RNR							
5 -	5.31(1.26)	-0.21(.05)	1.17(.23)	-0.94(.25)	-1.04(.08)	8.07	7.48
0	4.69(1.10)	0.00(.03)	1.03(.24)	0.02(.12)	0.01(.09)	8.03	7.88
+	4.38(1.32)	0.21(.05)	0.96(.29)	0.96(.23)	1.05(.10)	7.96	7.95
HNR							
6 -	5.19(0.81)	-0.21(.05)	1.14(.18)	-0.96(.21)	-1.06(.13)	7.98	7.71
0	4.75(0.75)	-0.00(.02)	1.05(.17)	-0.01(.10)	-0.03(.10)	8.09	8.04
+	4.81(0.73)	0.23(.03)	1.06(.16)	1.05(.13)	1.09(.11)	7.84	7.78

Because of calibration errors the r_{01} in the Rasch model may differ somewhat from the expected scores in OPLM at the corresponding θ_{01}. Table 5 shows that this effect of deviant calibration is most remarkable for condition 5 (RNR) at the negative decision level, where there is almost a difference of 0.6 in the expected scores under OPLM and the Rasch model. This deviance does not reappear in condition 6 (HNR) with the more homogeneous item bank, where Rasch calibration is expected to be more satisfactory.

For the discussion of the misclassification functions we will focus on decision level $\theta_{03} = 0.22$ and the long test of 64 items. Conspicuous deviant properties of the other decision levels and test length will be given additional attention.

The graph of the misclassification function of test condition 1 (ODW), the optimal test condition with calibration by OPLM, item selection on the basis of

the information at the decision level θ_{03}, and using weighted scores, is shown in Figure 2.

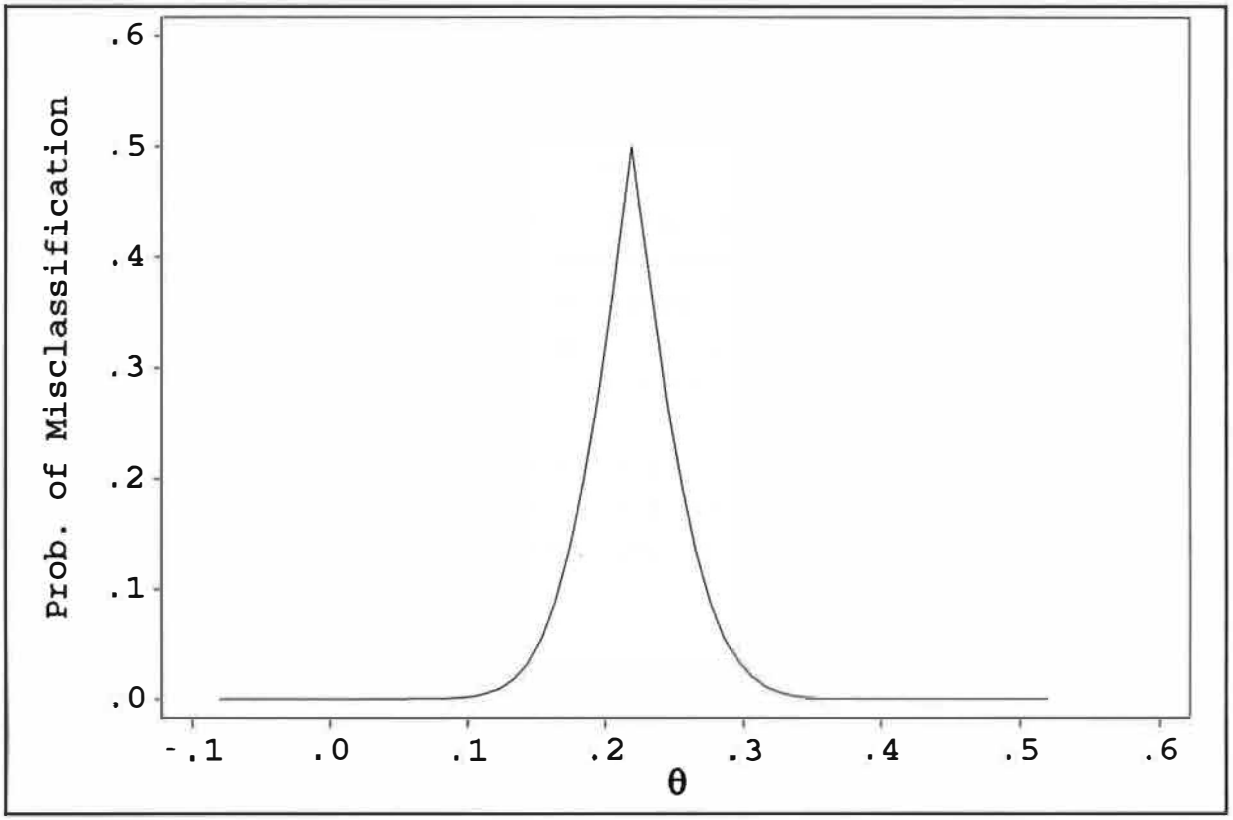

Figure 2. Misclassification function for test condition 1 (ODW) and the long test (64 items) at decision level θ_{03} = 0.22.

Figure 2 clearly shows that the probability of misclassification at values of θ very near θ_{03} is about 50%, as is reasonably to be expected. However, especially for the two regions at some distance from θ_{03} the probabilities of misclassification are interesting. and there we have to look for relevant differences with the five nonoptimal test conditions. These differences are graphically displayed in Figure 3. The number tags represent the test conditions 2 through 6. The right top at θ_{rt} = 0.26 reaches 8.3%, indicating that 8.3% more students with ability around θ_{rt} may undeservedly pass the test if condition 5 (RNR) applies as compared to condition 1 (ODW). The optimal test condition passes only 16.5% incorrectly against 24.8% for condition 5 (RNR) at θ_{rt}, at 4.5 x 0.04 = 0.18 standit from the decision level.

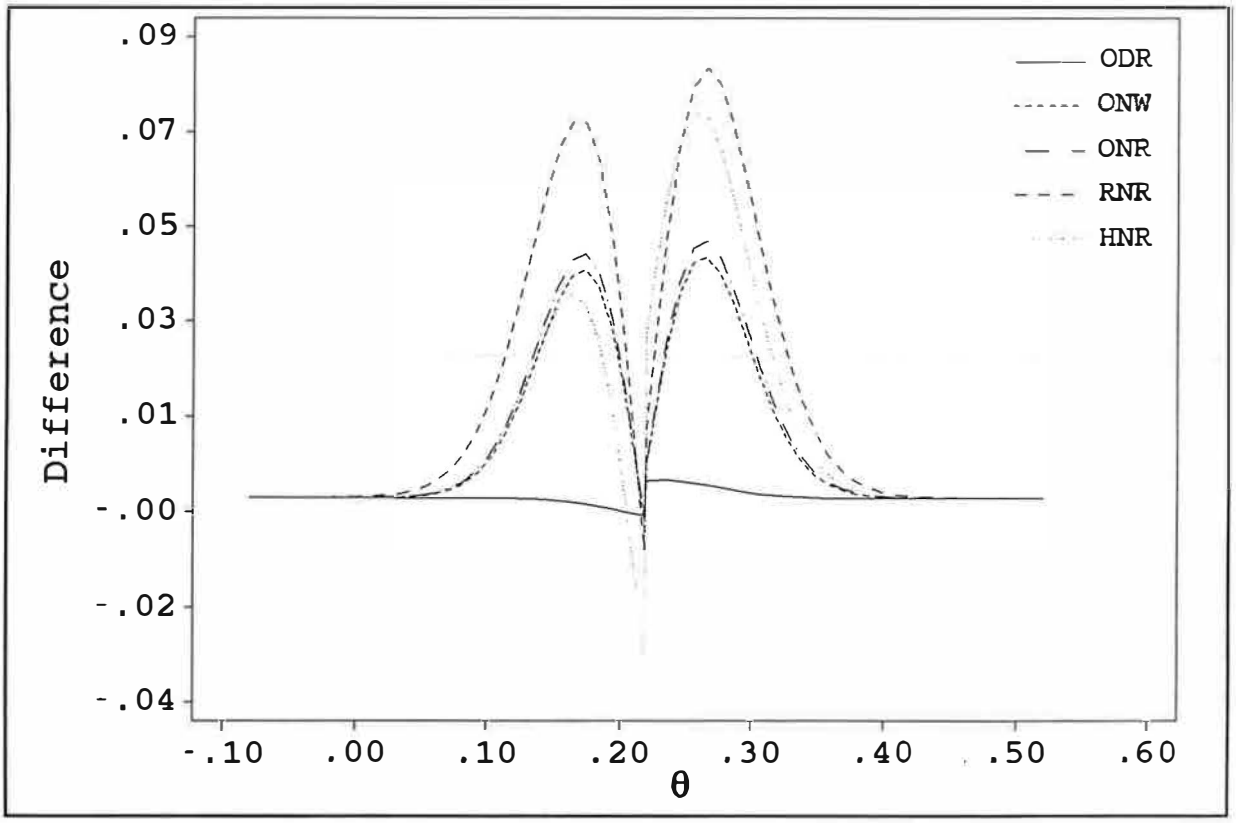

Figure 3. The differences between the misclassification functions of the nonoptimal conditions 2 through 6 (ODR through HNR) with the optimal condition 1 (ODW) for the long test at decision level $\theta_{03} = 0.22$.

A conspicuous property of the short test conditions at the same positive decision level is the asymmetry of the difference curve for condition 2, although the absolute percentages are relatively small (±2%). This is shown in Figure 4.

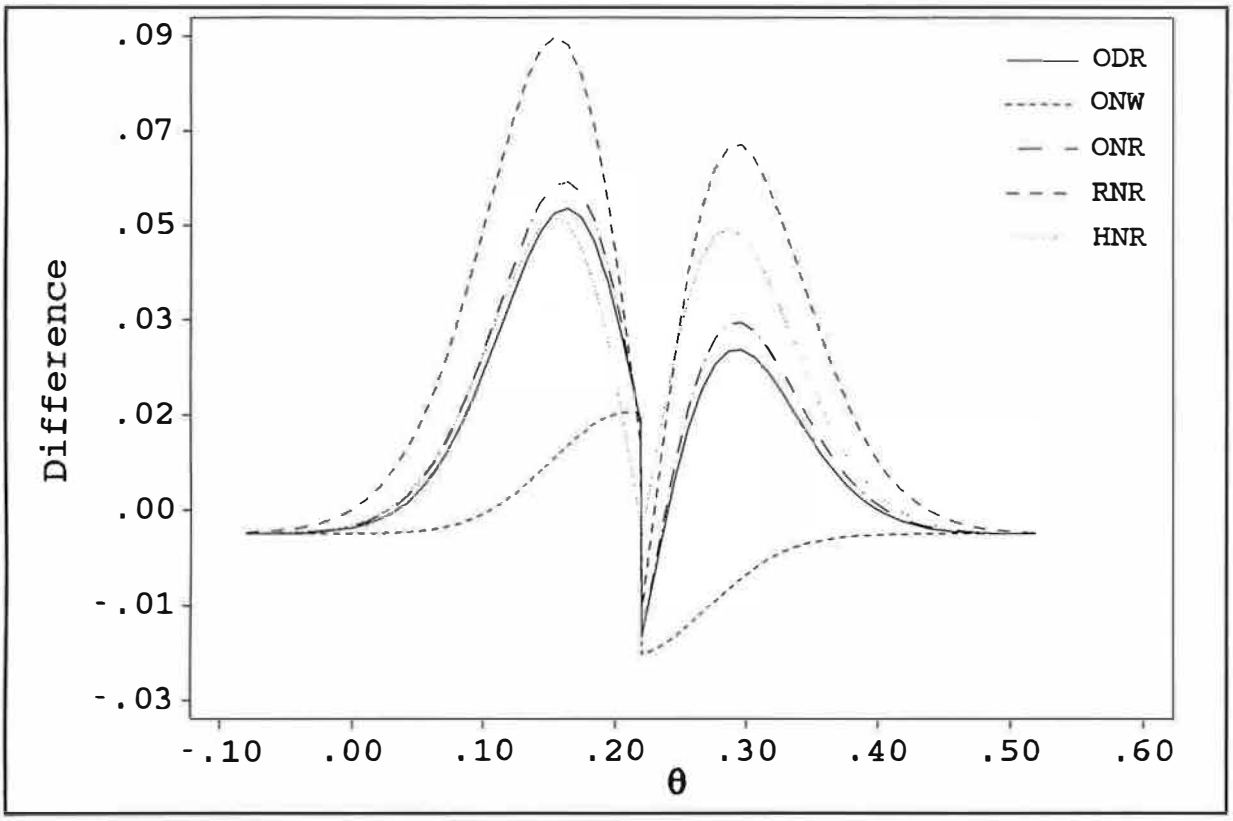

Figure 4. The difference functions for the short tests at the positive decision level. Note the asymmetries of the difference functions.

Because the misclassification function of condition 2 (ODR) appears almost perfectly symmetrical, the asymmetry of the difference curve proves to be caused by the asymmetry of the misclassification function of condition 1 (ODW). This asymmetry can be interpreted as the result of somewhat positively skewed conditional distributions of the weighted score for values of θ around θ_{03}. For instance, at $\theta = \theta_{03}$ about 52% of the conditional weighted score distribution falls below its mean $w_{03} = 57.69$. This means that for a value of θ just below θ_{03} 48% is misclassified, while for a value of θ just above θ_{03} 52% is misclassified. However, if we look at the results for the long test in Figure 3, the asymmetry is reversed, and less conspicuous as well. Considering that the long test is just four times the same 16 items, that were taken two times for the short test, these asymmetries must be considered as rather arbitrary effects. The reason is probably the rather hectic, though not irregular, shape of the conditional score distributions. As an example, the conditional weighted

score distribution of the 32 item test given that $\theta = 0.22$ is shown in Figure 5. The large differences in probability between closely spaced weighted scores is caused by the fact that they may differ greatly in the number of response patterns that produce them. To illustrate this the probabilities x 100 and the number of possible response patterns x 10^{-6} of the scores 110 through 122, at the centre of the distribution are given in Table 6.

TABLE 6

The centre of the conditional weighted score distribution at θ_{03}

w	110	111	112	113	114	115	116	117	118	119	120	121	122
%	0.36	0.44	1.84	4.11	4.77	2.63	0.55	0.27	1.36	3.59	4.91	3.18	0.76
#pat	22	73	137	140	76	21	12	41	92	110	70	22	6

There are, for instance, 21900692 possible response patterns that give rise to weighted score 110 in a 32 item test with 24 discrimination indices equal to 6 and 8 equal to 7. The two rows are, of course, not exactly parallel, because the first is conditional on θ, but the numbers of possible patterns heavily influences the conditional score distribution.

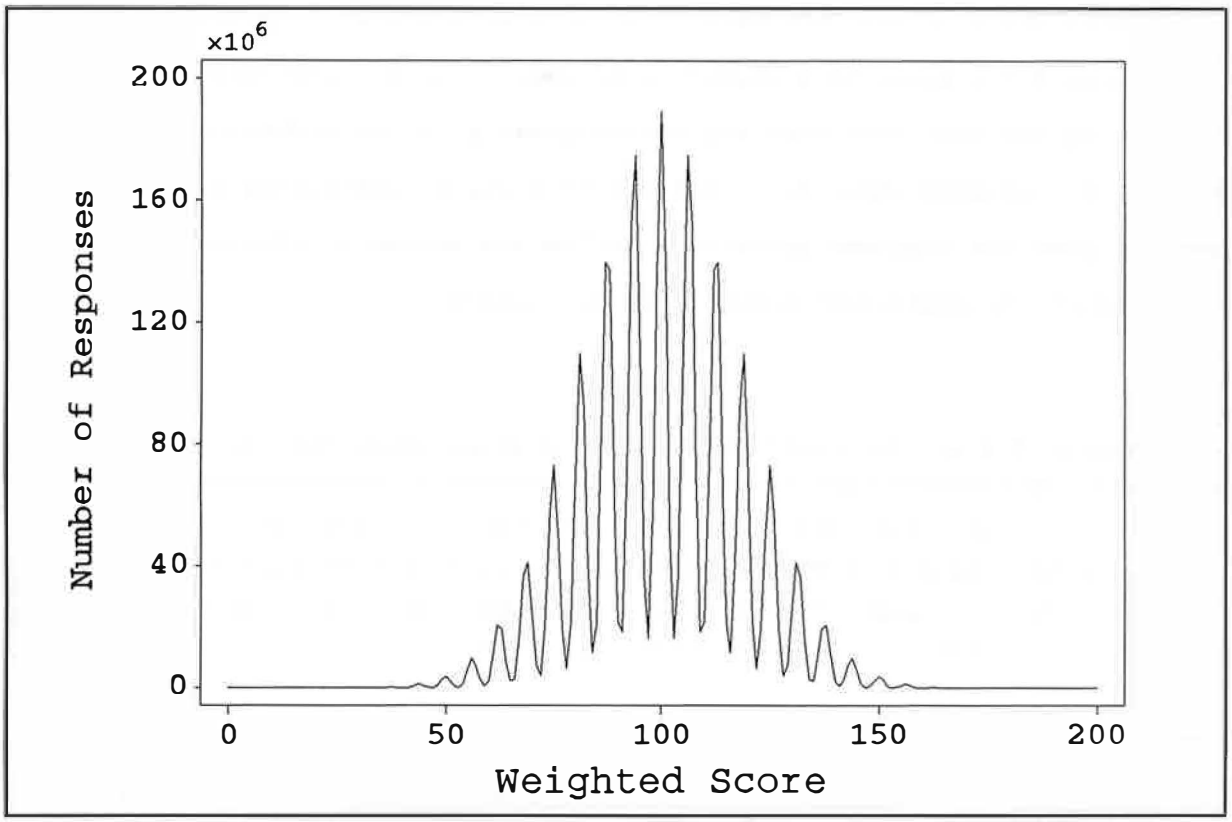

Figure 5. The number of response patterns per weighted score in a 32 item test with 24 discrimination indices equal to 6 and 8 equal to 7.

Worth noting further is that conditions 2 and 4 (ODR, ONR), with the raw score, incur only a minor loss in decision accuracy in comparison with their weighted score equivalents 1 and 3 (ODW, ONW) resp. For the long tests at θ_{rt} condition 2 (ODR) misclassifies only 0.3% more than condition 1 (ODW), and condition 4 (ONR) also only 0.3% more than condition 3 (ONW).

The decision accuracy is, however, appreciably diminished (>4%) by test construction without the use of discrimination indices. Of all the conditions, calibration and construction under the Rasch model of the complete item bank clearly performs worst. Selecting the more homogeneous subset yields a slight improvement, but does not match the OPLM conditions, not even those conditions where discrimination indices are neglected in test construction.

The other decision levels and test length, in principle show the same picture, except that the shift effect, the result of the hectic conditional weighted score distributions, may be reversed from left to right. A dramatic

shift effect can be observed for the two test lengths at the negative decision level $\theta_{01} = -0.22$. Here, the cause is not primarily to be found in slightly asymmetric conditional score distributions of condition 1 (ODW), but in the Rasch calibration errors that cause the deviation between expected raw scores in the RM and OPLM, as shown in Table 5. According to the OPLM-reality both expected scores around the decision level in conditions 5, and 6, (RNR, HNR) are appreciably less than their respective criterion scores. This results in a very low probability of misclassification for values of θ a little less than θ_{01}, and for values of θ a little greater than θ_{01} this must be paid for by a relatively high misclassification probability. For instance with the long 64 item test 40%, and 30% for conditions 5 and 6 resp. against only 15% for condition 1 at $\theta = 0.16$. See Figure 6 for the difference curves of the long tests at the negative decision levels. That the asymmetry of condition 1 (ODW) plays only a minor role here, is evidenced by the flat form of the difference curve for condition 2 (ODR). The important point here is that the effects of calibration errors in the Rasch model cannot be controled, and therefore add to the uncertainty about the quality of decisions made under the Rasch model. This uncertainty must be added to the reported measurement errors that assume that the model fits the data well.

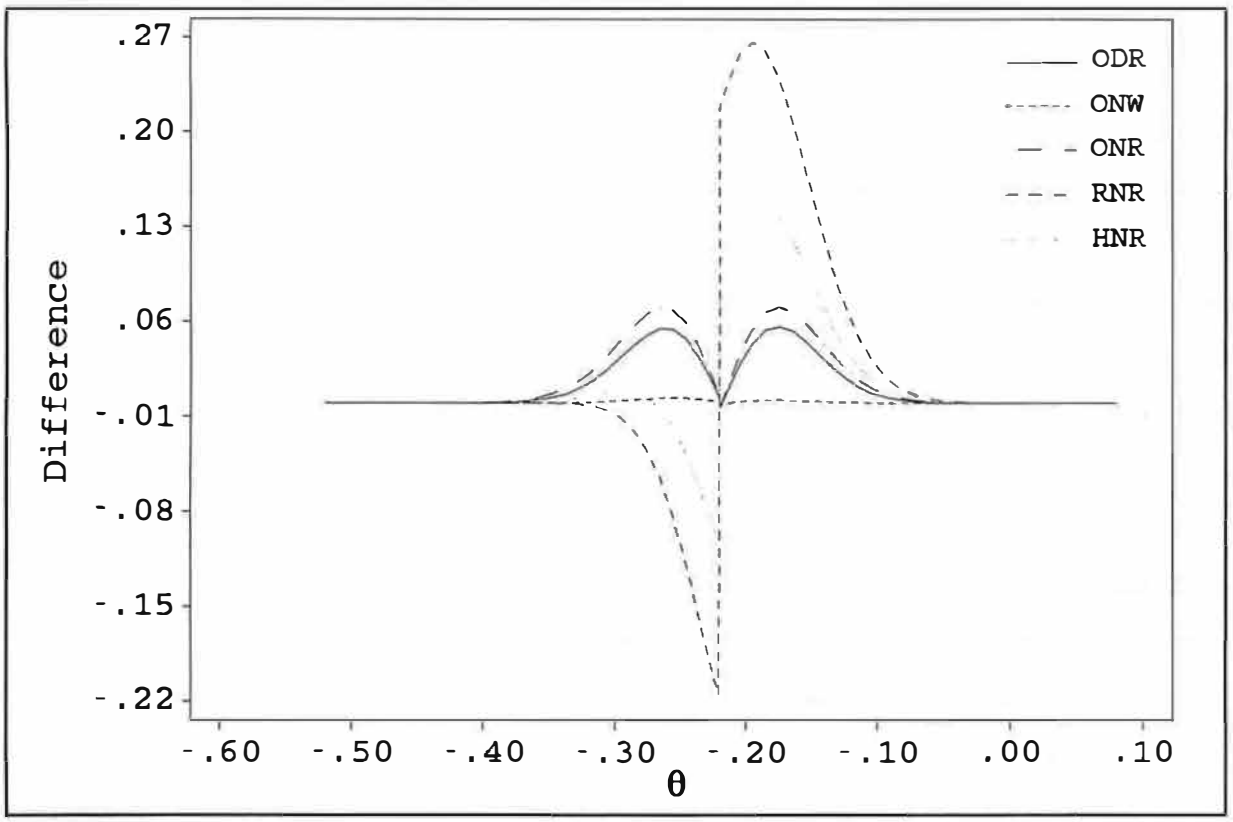

Figure 6. Difference functions for the long test conditions at the negative decision level. The large asymmetry for the Rasch conditions is the result of calibration errors caused by the unappropriateness of the Rasch model for the given data.

Conclusion

The above analysis shows that the choice of IRT model may influence the measurement accuracy of a test. Particularly in condition 5 (RNR), where the Rasch model was used most inappropriately, without the prior selection of a more homogeneous item subset, an appreciably larger percentage of students passes or fails the exam unwarranted, than if OPLM was used as the IRT model. The choice of the model by itself determines to a large extent the amount of injustice that is accepted. But also condition 6 (HNR), where the Rasch model is more appropriately used, at the cost of a fair amount of precious items, shows only a slight improvement over condition 5, and still produces an unnecessary, because easily avoidable, percentage of incorrect and unjust decisions that may harm individual careers or society.

If the use of weighted scores would pose a problem or be prone to errors, as in hand scoring, only a minor loss in accuracy is incurred by the use of raw scores, provided that the test is constructed with full use of the OPLM calibration results. However, even this minor accuracy loss may have a substantial effect on the value of the classical reliability coefficient. For a moderately reliable test of 47 items we witnessed an increase from 0.61 to 0.68, and increases of about 0.04 are common for moderate reliabilities.

A disadvantage of optimal test construction is that preferably highly discriminating items are selected. This may quickly damage the value of an item bank by too frequent use of the same small subset of items. Therefore, it is recommended when an item bank is used for decisions of various importance, to restrict optimal test design for the really important decisions. Test construction for less consequential decisions should then be carried out with disregard of the discrimination indices of the items. It is expected that the loss of accuracy here is more than compensated for by the increased quality of decisive tests. OPLM calibration enables one to make deliberate choices in the discussed respects, which certainly must be regarded as a valuable reward for choosing a more complicated model than the Rasch model.

References

Birnbaum, A. (1968). Contributions in: F.M. Lord and M.R. Novick, *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Staphorsius, G., Verhelst, N.D. and Kleintjes, F.G.M. (1991). *Reading Comprehension Tests at the Elementary School Level: Technical Reference*. Cito, Arnhem.

Verhelst, N.D., Glas, C.A.W. and Verstralen, H. F.M. (1991). OPLM: *A compunter program and manual*. Arnhem: Cito.

Verstralen, H.H.F.M. and Verhelst, N.D. (1991). *The sample strategy of a test information function in computerized test design*. Arnhem: Cito.

Recent Measurement and Research Department Reports:

91-1 N.D. Verhelst & N.H. Veldhuijzen. A New Algorithm For Computing Elementary Symmetric Functions And Their First And Second Derivatives.

91-2 C.A.W. Glas. Testing Rasch Models For Polytomous Items: With An Example Concerning Detection Of Item Bias.

91-3 C.A.W. Glas & N.D. Verhelst. Using The Rasch Model For Dichotomous Data For Analyzing Polytomous Responses.

91-4 N.D. Verhelst & C.A.W. Glas. A Dynamic Generalization Of The Rasch Model.

91-5 N.D. Verhelst & H.H.F.M. Verstralen. The Partial Credit Model With Non-Sequential Solution Strategies.

91-6 H.H.F.M. Verstralen & N.D. Verhelst. The Sample Strategy Of A Test Information Function In Computerized Test Design.