Measurement and Research Department Reports 97-4

Comparison of Test Administration Procedures for Placement Decisions in a Mathematics Course

G.J.J.M. Straetmans
T.J.H.M. Eggen

Cito

Measurement and Research Department Reports 97-4

Comparison of Test Administration Procedures for Placement Decisions in a Mathematics Course

G.J.J.M. Straetmans
T.J.H.M. Eggen

Cito
Arnhem, 1997

Abstract

In this study, three different test administration procedures for making placement decisions in adult education were compared: a paper-based test (PBT), a computer-based test (CBT), and a computerized adaptive test (CAT). All tests were prepared from an item response theory calibrated item bank. The subjects were 90 volunteer students from three adult education schools. They were randomly assigned to one of six experimental groups to take two tests which differed in mode of administration. The results indicate that test performance was not differentially affected by the mode of administration and that the CAT always yielded more accurate ability estimates than the two other test administration procedures. The CAT was also found to be capable of making placement decisions with a test that was on average 24% shorter.

Introduction

The goal of adult basic education in the Netherlands is to provide adults with knowledge and abilities that are indispensable for functioning satisfactorily as an individual and as a member of society. One of the courses in the programme is a mathematics course at three different levels of difficulty. A placement test is used to assign prospective students to a particular course level. As the candidates' abilities vary greatly, the written placement test currently being used is a two-stage test (Lord, 1971). In the first stage, all candidates take a routing test of 15 items with an average difficulty that corresponds to the average ability of the prospective students. Depending on their scores on the routing test, the candidates take one of three different measurement tests of 10 items each, varying in difficulty and adapted to the result of the routing test. There are several drawbacks to this procedure:

- the accuracy of measurement is lower due to misroutings. Because only one branching decision is used, possible misroutings cannot be corrected (Weiss, 1974);
- test administration is laborious because of the extra evaluation that has to take place after the first stage;
- maintaining confidentiality about the contents of the test is difficult owing to the flexible intake procedure which is a characteristic feature of adult basic education. This can easily lead to misclassifications (assignment of prospective students to a course level for which they lack ability or for which they are over-qualified).

Computerized adaptive testing might offer a solution to the problems mentioned:

- computerized adaptive tests have as many branching decisions as items in the test. Erroneously branching to items that are too easy (the response was wrong by mistake) or too difficult (the candidate made a good guess) is likely to be rectified by the remaining items in the test;
- computerized test administration offers the advantage of immediate test scoring and feedback. (Didactic measures can be taken immediately after the test.);
- maintaining confidentiality about the contents of the test is less of a problem because each testee takes a different test.

These features of computerized adaptive testing are very interesting, the more so since all schools for adult basic education will soon have well-equiped computer rooms at their disposal. The question is whether a computerized adaptive test version can have the psychometric quality (in terms of accuracy of measurement and efficiency) of the

written placement test. In a review of research Weiss and Schleisman (1994) come to the following conclusion: "Adaptive tests not only provide increases in measurement efficiency, but also improved measurement precision." However, these findings cannot be generalized to each application of computerized adaptive testing, because too little is known about the relationship between the quality and structure of an item bank, the functioning of the test algorithm, and the characteristics of the testees.

For the present test application, the psychometric superiority of a computerized adaptive test version over a written version of the placement test has already been demonstrated in a simulation study. Among others it was concluded that a CAT would require between 55% and 75% of the number of items of the paper-and-pencil version (Eggen & Straetmans, 1996). However, a new study under real placement test conditions is necessary for the following reasons:

- many students in adult basic education have little or no experience with computers. The occurrence of operating problems and/or computer anxiety may interfere with test performance;
- a computerized adaptive test restricts the freedom of testees to take the test in their own way. Browsing through the test and deferring the response to more difficult items until the end of the test, for instance, are simply not possible. Several researchers have found evidence that test performance increases when testees have more control over the test procedure (Roos, Plake & Wise, 1992; Vispoel & Coffman, 1994; Wise, Plake, Johnson & Roos, 1992);
- it is possible that the interpretation and accuracy of test scores is affected by the fact that calibration of the item bank was based on paper-and-pencil administrations.

Research Questions

The present study was carried out to answer the following questions:

1 Are ability estimates and placement decisions affected by test administration procedures?

2 Does the CAT yield more accurate measurements of mathematics ability than the paper-based version of the placement test?

3 Is the CAT more efficient than the paper-based version of the placement test?

Method

Subjects

Ninety subjects from three schools for Basic Education volunteered to participate in the study. In particular, these schools were invited to participate because of their well-equiped computer rooms (at least 10 PCs (IBM or clone) with a 386 cpu or higher, Windows 3.xx, and connected in a LAN).

Design

Although the purpose of this investigation was to decide whether a computerized adaptive test is preferable as a placement test to a paper-based test, *three* different administration procedures for a mathematics placement test were compared: paper-based test (PBT), computer-based test (CBT), and computerized adaptive test (CAT). The CBT was included as it might be helpful in explaining possible significant differences between the mean ability estimates for PBT and CAT.

The study was carried out according to the experimental design in Table 1. Students were randomly assigned to one of six experimental groups.

Table 1

Research Design

Group	First test	Second test	Number of subjects
1	PBT	CBT	14
2	CBT	PBT	15
3	PBT	CAT	16
4	CAT	PBT	15
5	CBT	CAT	15
6	CAT	CBT	15

Therefore, in the experiment, individual students were randomly assigned to two out of three conditions (PBT, CBT, and CAT).

Instruments

Three different forms of the placement test were developed. All tests were prepared from an item response theory (IRT) calibrated item bank of 250 items (mainly short answer questions that can be scored dichotomously). The items were calibrated by

means of the OPLM model (Verhelst & Glas, 1995). The basic equation in this model is:

$$p_i(\theta) = P(X_i = 1 \mid \theta) = \frac{\exp a_i(\theta - \beta_i)}{1 + \exp a_i(\theta - \beta_i)}.$$

The response to an item X_i is either correct (1) or incorrect (0). The probability of scoring an item correctly is an increasing function of latent ability θ and depends on two item characteristics: the item difficulty β_i, and the discriminatory power of an item a_i. A weighted maximum likelihood estimate was used to estimate the ability of the students (Warm, 1989).

The PBT and the CBT are identical except for the mode of administration. Both PBT and CBT are two-stage tests. The third administration procedure is a CAT. The CAT is different from PBT and CBT in that the test is constructed on-line. Each time the candidate has responded to an item, the test algorithm selects a new item that provides a maximum level of item information at the candidate's currently estimated ability. The test algorithm has the following specifications:

- in order to reduce possible negative effects of test anxiety, the first three items are selected at random from a subset of relatively easy items;
- the selection of the remaining items is based on the principles of maximum information at the current ability estimate (Weiss & Kingsbury, 1984);
- a fixed test length of 25 items is used to facilitate comparisons between the administration procedures.

The function of the mathematics placement test is to assign prospective students to one of three course levels: level 1 (beginners), level 2 (intermediate) and level 3 (advanced). For this purpose, two cut-off scores were defined on the ability scale that resulted from the calibration. The exact positions on the scale were determined as follows:

- content specialists defined subsets of items by labeling the items in the item bank as level 1, 2, or 3 items;
- the mean difficulty was defined for each subset of items;
- using the basic equation of the OPLM model, the lower and higher cutting points were defined by calculating the abilities that give a probability of success of at least 0.7, given the mean difficulties of the subsets of level 1 and level 2 items, respectively. This resulted in $\theta_1 = -0.13$ (lower cutting point) and $\theta_2 = 0.33$ (higher cutting point).

Procedure

At the beginning of each testing session, the subjects were instructed on how to operate the computerized tests. They also received a written summary of the instructions which could function as a 'job aid' while taking a computerized test. Next, the subjects took the first test. After completing the first test, the subjects were asked to fill out a short questionnaire about their feelings and experiences with regard to specific characteristics of this test. After a short break, the subjects took the next test. Again, a short questionnaire was administered after the subjects had finished the second test.

Analysis and Results

Person fit analysis

The items in the item bank from which the PBT, the CBT, and all versions of the CAT had been prepared, were calibrated from data collected in a paper-and-pencil administration. Mills and Stocking (1996) cast doubt on the appropriateness of item parameter estimates when the calibration medium differs from the test medium. Their doubts are based on a study by Divgi and Stoloff (1986), who have shown that item response functions of items administered in a CAT differed from item response functions of the same items administered in a PBT. In the present study, justification for the use of paper-based item parameter estimates in a CBT and a CAT was obtained by performing a person fit analysis. A person fit analysis indicates how likely an observed response pattern is, given a person's test score or estimated ability and presuming that the paper-based item parameter estimates are valid. In this study the caution index 'Zeta' (Tatsuoka, 1984) was computed for each of the 180 item response patterns. Zeta can be interpreted as a standard normal variate, which means that indices ≥ 1.96 or ≤ -1.96 indicate a significant deviation ($p \leq .05$) of the observed item response pattern in relation to the expected item response pattern. Zeta tends to be positive if a respondent has too many correct answers on difficult items and too few on easy items. A negative Zeta results from respondents obtaining too many correct scores on easy items and too few correct scores on difficult items. Table 2 summarizes the outcomes of the person fit analysis. Significant deviating response patterns were observed in only 7 out of 180 cases. This means that there are no indications that the paper-based item calibrations cannot be used for administering and scoring computer-based tests.

Table 2

Person Fit Analysis Results

Test procedure	Number of tests	Mean Zeta	SD	Number of negative deviations	Number of positive deviations
PBT	60	-0.020	0.782	0	1
CBT	59	0.491	1.126	0	5
CAT	61	-0.215	0.798	1	0

Descriptive statistics

Table 3 contains descriptive statistics for the ability estimates and their standard errors, which were calculated over the following arrangements of test scores:

overall: all 180 test scores taken together;

1st test/2nd test: the test scores of the subjects who took their first/second test;

PBT/CBT/CAT: the test scores of the subjects who took a PBT/CBT/CAT;

PBT-1/CBT-1/CAT-1: the test scores of the subjects who took a PBT/CBT/CAT as their first test;

PBT-2/CBT-2/CAT-2: the test scores of the subjects who took a PBT/CBT/CAT as their second test.

Table 3 shows some striking results which deserve particular attention. In the first place the difference in mean theta between the first test and the second test. The magnitude of this difference raises the question whether this should be interpreted as an order effect. If that is the case, it does not make sense to evaluate decision consistency as partial evidence for the absence of a test administration procedure effect (research question 1).

Another aspect deserving special attention is the low mean theta for CBT-1 (0.079). The relatively low standard deviation for this mean (0.300) suggests that the low average ability cannot be accounted for by a few very low-ability and/or very high-ability subjects who took the CBT as their first test. The crucial question is whether the low mean theta for CBT-1 has to be interpreted as an indication of a test administration procedure effect. Such an effect would make this study inconclusive. The possibility of both the order effect and the test administration procedure effect was investigated and will be dealt with in detail in the next section.

Table 3

Descriptive Statistics for Theta and Standard Error

		Theta				Standard error			
Arr	n	mean	SD	low	high	mean	SD	low	high
overall	180	0.264	0.397	-1.343	1.612	0.119	0.036	0.096	0.463
1st test	90	0.241	0.395	-1.343	1.386	0.118	0.039	0.097	0 463
2nd test	90	0.288	0.399	-0.894	1.612	0.120	0.032	0.096	0.324
PBT	60	0.289	0.406	-0.447	1.386	0.128	0.049	0.110	0.463
CBT	59	0.198	0.368	-0.894	0.977	0.121	0.018	0.110	0.209
CAT	61	0.305	0.412	-1.343	1.612	0.109	0.030	0.096	0.324
PBT-1	30	0.296	0.409	-0.344	1.386	0.132	0.064	0.110	0.463
CBT-1	30	0.079	0.300	-0.463	0.667	0.116	0.005	0.110	0.128
CAT-1	30	0.346	0.424	-1.343	1.004	0.106	0.016	0.097	0.183
PBT-2	30	0.281	0.409	-0.447	1.136	0.124	0.027	0.110	0.257
CBT-2	29	0.320	0.396	-0.894	0.977	0.125	0.025	0.110	0.209
CAT-2	31	0.265	0.403	-0.410	1.612	0.112	0.039	0.096	0.324

Note. Arr = Arrangement of test scores. Low = low extreme. High = high extreme.

The standard errors confirm what was expected, namely, that abilities can be estimated more accurately with CAT than with CBT or PBT. Whether these differences are meaningful and what the consequences are in terms of efficiency will be discussed below.

Is test performance affected by the test administration procedure?

Although the main purpose of the experiment was to investigate whether there are differences between the three test administration procedures with respect to the estimated abilities, the descriptive statistics in Table 3 reveal a possible order effect of the administration procedure which could interfere with this. One way to unravel these effects is by using a multilevel approach in the data analysis. First, this approach is explained, and then the results of the analysis are presented.

The data from the experiment may be considered as gathered in a repeated measurement design (three repeated measures per subject) from which one measure is missing at random. However, traditional methods for analyzing this kind of data, such as multivariate analysis of variance (manova), require complete data or the use of imputation procedures for the missing data. A flexible way out of this problem, indicated by Goldstein (1995), among others, is to consider the repeated measures as

multilevel data, or, in this application, two-level data. The measurements in each of the three conditions (short for test administration procedures) define the first level, and the subjects define the second level.

Snijders and Maas (1996) show how the computer program MLn (Rasbach & Woodhouse, 1995) can be used to obtain statistically sound estimates and tests of the manova model with incomplete data. Their approach was applied to the data of the experiment, supplementing it with some general modelling facilities and statistical tests of the MLn program. The models used will now be described, starting with the most general model, in which both effects of the administration procedures as an order effect are estimated. After that, more restricted models are formulated, which provide the opportunity to test relevant hypotheses.

The dependent variable used in the models is the vector: $\hat{\theta}_j = (\hat{\theta}_{1j}, \hat{\theta}_{2j}, \hat{\theta}_{3j})$, with $\hat{\theta}_{ij}$ being the estimate of the ability of subject j in condition i; $i = 1, 2, 3$ for PBT, CBT, and CAT, respectively. The mean abilities in these conditions are indicated by β_1, β_2, and β_3, respectively. The parameter for the order effect, that is, the difference in ability between the second and the first administration, is denoted by δ. The general model can then be specified by defining dummy variables. Three dummy variables $x_{1ij}, x_{2ij}, x_{3ij}$, one for each condition: $x_{iij} = 1$ and $x_{hij} = 0$ for $h \neq i$. This means, for example, that $x_{1ij} = 1$ only if subject j is measured in condition 1 (PBT). A fourth dummy variable indicates whether an administration condition is given first or second to a subject: $x_{4ij} = 0$ if the first administration condition of subject j is i, and $x_{4ij} = 1$ if the second administration condition of subject j is i.
The general model is then given by

$$\text{Model 1: } \hat{\theta}_{ij} = \beta_i + \delta x_{4ij} + U_{ij} = \sum_{h=1}^{3} \beta_h x_{hij} + \delta x_{4ij} + \sum_{h=1}^{3} U_{hj} x_{hij}, \quad i = 1, 2, 3; \; j = 1, \ldots, n.$$

In this model and the restricted models which follow, it is assumed that $U_j = (U_{1j}, U_{2j}, U_{3j})$ is multivariate normal distributed with mean $\mathbf{0} = (0, 0, 0)$, and covariance matrix given by $var(U_{ij}) = \sigma^2_{u,i}$, $i = 1, 2, 3$ and $cov(U_{ji}, U_{jk}) = \sigma_{u,ik}$, $i \neq k$. Thus the two-level model is empty at level 1 and, at level 2, there are fixed parameters for the three condition dummies and for the order dummy. As the random part of the model, we have correlated slopes of the condition dummies: the means and the covariance matrix of the three dependent variables.

In the sequel likelihood ratio tests will be used to answer the research questions. The general idea is that the likelihood of a restricted model, which expresses a hypothesis, is compared to the likelihood of the general model (model 1).

In order to investigate the occurrence of a significant order effect, δ, the following restricted model was formulated:

$$\text{Model 2: } \hat{\theta}_{ij} = \sum_{h=1}^{3} \beta_h x_{hij} + \sum_{h=1}^{3} U_{hj} x_{hij}, \; i = 1,2,3; \; j = 1,\ldots,n.$$

To test the equality of the means of the conditions, that is, $\beta_1 = \beta_2 = \beta_3 = \mu$, indicating that the means of the conditions are all equal to a general mean (Snijders & Maas, 1996), another restriction of model 1 was formulated:

$$\text{Model 3: } \hat{\theta}_{ij} = \mu + \delta x_{4ij} + \sum_{h=1}^{3} U_{hj} x_{hij}, \; i = 1,2,3; \; j = 1,\ldots,n.$$

Finally, we need to consider

$$\text{Model 4: } \hat{\theta}_{ij} = \mu + \sum_{h=1}^{3} U_{hj} x_{hij}, \; i = 1,2,3; \; j = 1,\ldots,n,$$

expressing neither differences in condition means nor an order effect. It will be clear that this model is a restriction of both model 2 and model 3 and by that of model 1.

The results of the parameter estimates of these four models are given in Table 4. Table 4 shows that there are hardly any differences between the random parts of the four models. The variance in the CBT ability estimates ($\sigma^2_{U,2}$) is somewhat lower (about .14) than in the other conditions (about .16). The correlations between the abilities in the administration conditions are easily computed from Table 4 by dividing the covariance by the square root of the variances. As expected, all correlations are high: between PBT and CBT about .94, between PBT and CAT .90, and between CAT and CBT .86.

Of primary interest are the results of the fixed part of the estimated models. It is seen in model 2 that the mean abilities for the administration conditions PBT, CBT, and CAT are .287, .252 and .252 respectively. Compared to the descriptive results in Table 3, reporting means of respectively .289, .198, and .305, the differences between the means have become smaller. This can be explained by the fact that the values for the means in Table 3 would have been the estimates of the means if a clearly wrong model, assuming equality of the variances of the abilities in the three conditions and

independence between all observations, would have been applied to the data. The question whether there are still significant differences between the means and whether there is a significant order effect will be answered now.

Table 4

Parameter Estimates, Standard Errors, and Likelihoods of the Multilevel Models

Model	1		2		3		4	
Fixed								
$\beta_1(se)$	.258	(.044)	.287	(.043)				
$\beta_2(se)$	.224	(.041)	.252	(.041)				
$\beta_3(se)$	.224	(.046)	.252	(.044)				
$\mu(se)$					.229	(.041)	.261	(.040)
$\delta(se)$	.055	(.017)			.056	(.017)		
Random								
$\sigma^2_{U,1}(se)$	.162	(.025)	.160	(.025)	.162	(.025)	.160	(.025)
$\sigma_{U,21}(se)$	.140	(.022)	.141	(.022)	.142	(.022)	.144	(.023)
$\sigma^2_{U,2}(se)$	.135	(.021)	.143	(.023)	.139	(.022)	.148	(.023)
$\sigma_{U,31}(se)$	.150	(.024)	.147	(.024)	.147	(.024)	.145	(.023)
$\sigma_{U,32}(se)$	.132	(.022)	.132	(.023)	.133	(.022)	.133	(.023)
$\sigma^2_{U,3}(se)$	.168	(.027)	.164	(.026)	.166	(.027)	.161	(.026)
-2*ln (lik)	12.836		22.204		15.684		24.917	

The hypothesis that $\delta = 0$, that is, there is no order effect, can be tested in two ways. The first is by inspecting the value of the estimate of the order effect parameter δ in model 1, .055, which is, compared to its standard error, significantly different from 0. The second way is to perform a likelihood ratio test to test the more restricted model 2 against model 1. The χ^2 statistic with 1 degree of freedom is $\chi^2_1 = 22.204 - 12.836 = 9.368$, with a probability of $p = .002$. This means that the hypothesis (model 2, $\delta = 0$) is rejected in favor of model 1.

Both tests indicate that there is a significant order effect in the experiment, which also means, as mentioned before, that it does not make sense to evaluate the consistency of the placement decision on the basis of the three administration procedures. Therefore, evidence for a possible test administration procedure effect can only be obtained from the estimated mean abilities.

The likelihood ratio test of the restricted model 3 against model 1 tests the hypothesis $\beta_1 = \beta_2 = \beta_3 = \mu$. The χ^2 statistic with 2 degrees of freedom is $\chi^2_2 = 2.848$ (15.684 - 12.836), which gives a probability of .24 under the hypothesis, indicating that hypotheses about the equality of all three means, model 3, cannot be rejected. The equality of the means were also testes pairwise with the MLn program, using a Wald test. The results for model 3 were $\beta_1 = \beta_2$, $\chi^2_1 = 2.55$, p = .11; $\beta_1 = \beta_3$, $\chi^2_1 = 1.75$, p = .19; and $\beta_2 = \beta_3$, $\chi^2_1 = 0.0$, p = 1.0. So none of the differences between the mean abilities of the administration procedures in this model are significant.

Model 2 is therefore rejected in favor of model 1, but model 3 cannot be rejected when tested against model 1. As it was not yet clear which model described our experimental data best, two other likelihood ratio tests were conducted.

The first is testing model 4 ($\beta_1 = \beta_2 = \beta_3$ and $\delta = 0$) against model 1, which gives $\chi^2_3 = 12.081$ and $p = .007$, indicating that model 4 has to be rejected in favor of model 1. And the second tests model 4 against model 3: $\chi^2_1 = 9.233$, $p = .002$, also rejecting model 4 in favor of model 3.

Models 3 and 1 were never rejected. From the hypothesis regarding model 3 against model 1, it must be concluded that model 3 cannot be rejected. But if these two models are compared with Akaike's (1974) criterion (for model 3: 31.684 and for model 1: 32.836), model 3 is to be preferred.

It can be concluded that the best model to describe the experimental data is model 3, which means that there is a significant order effect and that none of the differences between the mean abilities of the administration procedures are significant.

Does the CAT yield more accurate measurements?

The descriptive statistics of the standard errors of the ability estimates in Table 3 indicate that the highest measurement accuracy (after 25 items) is reached in the CAT administration condition. The meaningfulness of the observed differences in Table 3 will be reported in this section.

It may be possible to use the multilevel models 1 and 2 from the preceding section to estimate the means of the standard errors of the ability estimates in the three administration conditions and to test the significance of the differences (cf. Vispoel, Rocklin & Wang, 1994). However, inspection of the data showed that, even after logarithmic transformation, the assumed normality of the distributions is clearly not met (see Figure 1).

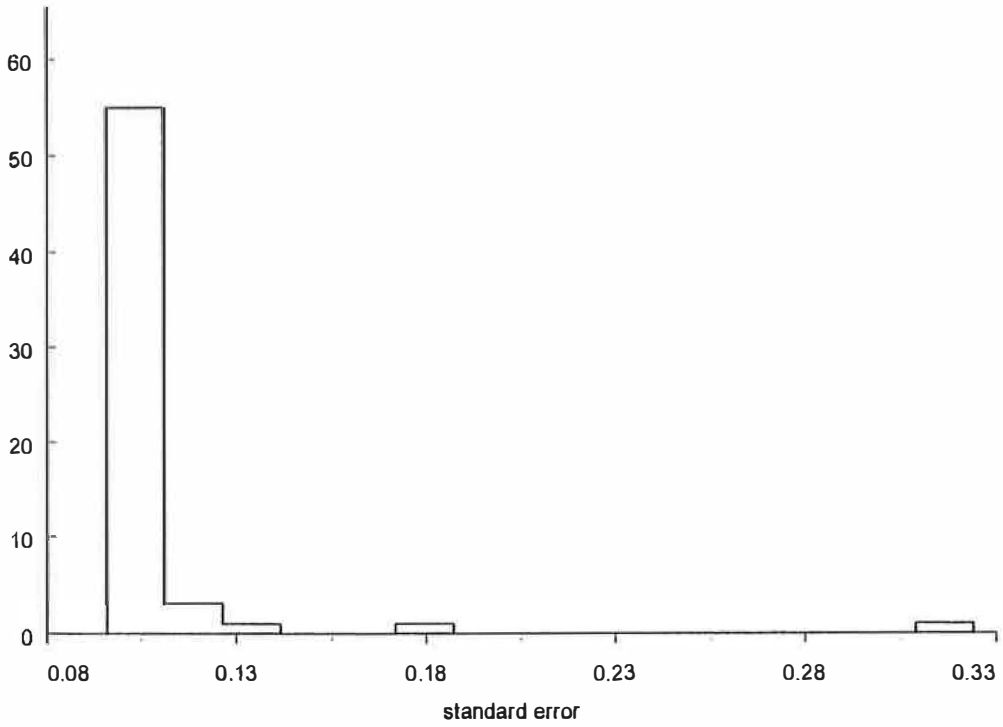

Figure 1

Histogram of the standard errors of ability estimates in the CAT

Typically, the distributions of the standard errors in each administration condition, but especially in the CAT, are very skewed: most of the students have been measured fairly precisely, tending towards a lower bound of the standard error and relatively few possibly able or unable students have higher or very high standard errors.

For this reason, the differences in the standard errors between the test administration procedures were tested by a nonparametric test, the Wilcoxon signed-rank test (Lehmann, 1975). The results are given in Table 5.

Table 5

Comparison of the standard errors of the ability estimates between the test administration procedures

Comparison	n	n^+	p-value
PBT-CBT	29	12	.796
PBT-CAT	31	31	.000
CBT-CAT	30	30	.000

In Table 5, n is the number of pairs of students involved in the comparison; n^+ is the number of positive differences in the standard errors between the first and the second mentioned test administration procedure; and the p-value for the Wilcoxon signed-rank test for no differences between the standard errors. The results are very clear. In the CAT, students are always measured more precisely than in any of the two other test administration procedures. Regarding the PBT-CBT comparison, one sees that somewhat more than half of the students are measured more precisely in the PBT condition than in the CBT condition, but this difference is not significant.

Is the CAT a more efficient procedure for taking placement decisions?

In the experiment, the students were administered 25 items in all three test administration procedures. When operational, the CAT will apply a stopping rule based on accuracy. Remembering that the function of the test is to assign students to one of three levels, it is clear that the efficiency of the CAT with a stopping rule compared to the full length CAT (and also to PBT and CBT) can be measured by the mean number of items, provided that the same decision was taken in the administration conditions that are to be compared. The CAT data were therefore reanalyzed applying the following stopping rule, which is described in more detail in Eggen and Straetmans (1996): after each item administered, say the k^{th}, a confidence interval for the examinee's true ability θ is constructed: $(\hat{\theta}_k - \gamma.se(\hat{\theta}_k)\,,\, \hat{\theta}_k + \gamma.se(\hat{\theta}_k))$, in which γ is a constant, determined by the required decision accuracy. As long as there is a cutting point, $\theta_1 = -.13$ or $\theta_2 = .33$, within the interval the next item (with a maximum of 25) is administered; if not, a decision is taken according to the following rules: assign to level 1 if the upper bound of the interval is smaller than θ_1, assign to level 3 if the lower bound of the interval is larger than θ_2, otherwise, assign to level 2.

Applying this stopping rule to the CAT data with $\gamma = 1.644$ (that is using 90% confidence interval), resulted in 60 out of 61 CAT administrations with exactly the same decision as the full length (25) CAT. The number of items required are given in Figure 2.

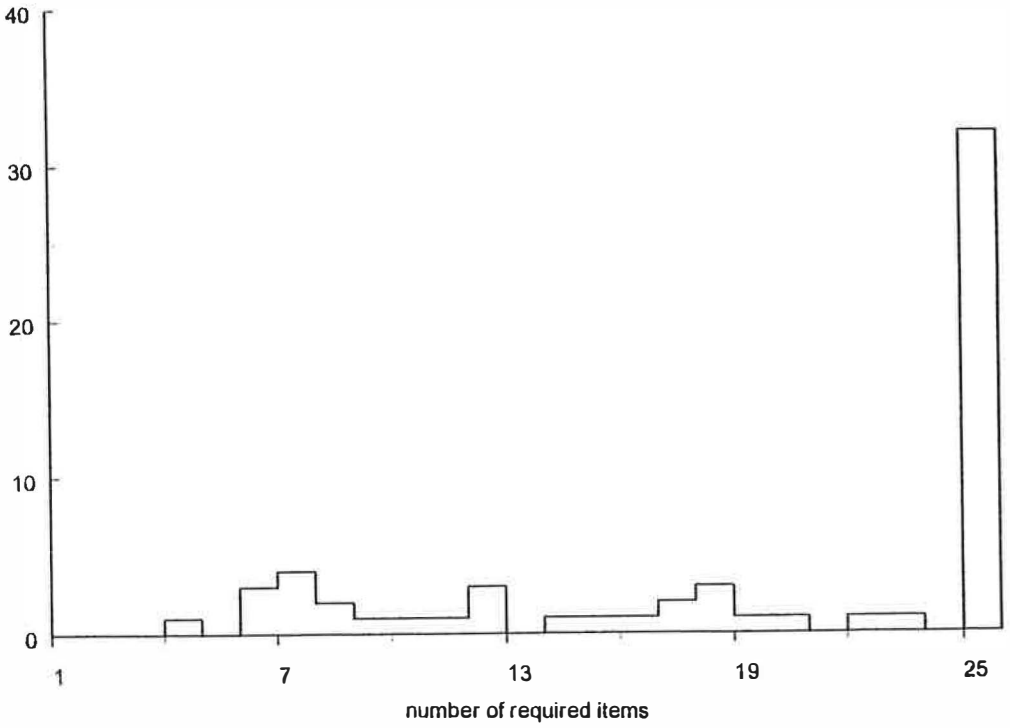

Figure 2

Histogram of the number of items required to make decisions about students

In about half of the administrations, the full test length was needed, but in the other half, a reduction in the number of required items could be obtained. The mean number of items needed was 19.1 (s.d. = 7.3), so that a reduction of the test length of about 24% on average is to be expected compared to the other administration conditions.

Discussion

The most important conclusion drawn in the previous section was that the mode of administration did not appear to be a factor in test performance. This was an important conclusion for the present study because it made it possible to answer research questions 2 and 3. For the future implementation of computerized adaptive testing, this conclusion was important because it is reassuring to know that in the transition stage to adaptive testing, both CAT and PBT can use the same scale.

The second conclusion confirmed what has been shown in recent years by many studies, namely, that adaptive tests have greater measurement precision than conventional tests. The greater measurement precision is in itself not very interesting. It is nothing more than a tool to realize more efficient testing by reducing the length of

the test. According to the guidelines for computerized adaptive test development (American Council on Education, 1995), a CAT uses about half the number of items required by traditional test models. Vispoel, Rocklin, and Wang (1994) even report reductions of almost 70%. The results of the present study could confirm the psychometric superiority of computerized adaptive testing over conventional testing, but the (to be expected) gain in efficiency was not as big, namely 24%. An explanation for this might be found in the specific purpose of the tests under study, namely making placement decisions. For that purpose, two cut off scores were defined on the ability scale. The distance on the ability scale between the lower (-0.13) and the higher cut off score (0.33) amounts to only 0.46 theta points, whereas the range of the ability scale is 2.95 theta points. Considering a mean standard error of the CAT of 0.109, a 90% confidence interval based on it, and more than 40% of the students who took a CAT having abilities between the lower and higher cut off score, it is not surprising that only half of the 61 CATs succeeded in making a placement decision with fewer than 25 items.

Measurement experts are very enthusiastic about the enhanced measurement performance of a CAT. However, it is not likely that teachers always share their feelings. For them, a reduction of test length will not be a decisive factor in considering the implementation of CAT. Other advantages of computerized adaptive testing are more appealing, especially when applied in adult education where flexible intake procedures raise problems of inefficiency in testing (e.g., one or a few students a day) and test security. Computerized adaptive testing offers a solution to both problems. Inefficiency is relieved by the fact that computers take care of laborious activities such as assembling, administering, and scoring tests. This makes it possible to test students on demand and just-in-time. Test security can be maintained because, in theory, each student takes a different test.

Perhaps the most important factor to be considered with regard to the implementation of computerized adaptive testing are the feelings and opinions of those who will be tested. The questionnaires that were filled out by the students after completing each test, showed that about 60% of them preferred taking a CBT or CAT to a PBT and that 20% had no preference. These are remarkable figures, the more so since 50% or more of the students indicated that they were more or less nervous while being tested.

The overall conclusion is that computerized adaptive testing can solve the problems which are characteristic of testing in adult education and that there are no objections to a possible implementation from a psychometric view.

In future research, attention will be paid to refinements of the use of statistical testing instead of statistical estimation as the computation procedure in the adaptive test algorithm. It has already been shown that this is a promising alternative leading to a larger gain in efficiency in the cases of classifying examinees into two categories (Spray & Reckase, 1996) and into three categories (Eggen & Straetmans, 1996).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automated Control, AC-19*, 716-723.

American Council on Education (1995). *Guidelines for computerized-adaptive test development and use in education.* Washington, DC: Author.

Divgi, D.R., & Stoloff, P.H. (1986). *Effect of the medium of administration on ASVAB item response curves* (Report No. 86-24). Alexandria VA: Center for Naval Analyses.

Eggen, T.J.H.M., & Straetmans, G.J.J.M. (1996). *Computerized adaptive testing for classifying examinees into three categories.* Manuscript submitted for publication.

Goldstein, H. (1995). *Multilevel statistical models*. 2nd edition. London: Edward Arnold.

Lehmann, E.L. (1975). *Nonparametrics: statistical methods based on ranks*. San Francisco: Holden-Day Inc.

Mills, C.N., & Stocking, M.L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*, 287-304.

Rasbach, J., & Woodhouse, G. (1995). *Multilevel Command Reference*. London: Institute of Education, University of London.

Roos, L.L., Plake, B.S., & Wise, S.L. (1992, April). *The effects of feedback in computerized adaptive and self-adapted tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Snijders, T.A.B., & Maas, C.J.M. (1996). Application: Using MLn for repeated measures with missing data. *Multilevel Modelling Newsletter, 8*, 7-10.

Spray, J.A., & Reckase, M.D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*, 405-414.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95-110.

Verhelst, N.D., & Glas, C.A.W. (1995). The one-parameter logistic model. In G.H. Fisher & I.W. Molenaar (Eds.), *Rasch models: their foundations, recent developments, and applications* (pp. 215-237).

Vispoel, W.P., & Coffman, D.D. (1994). Computerized-adaptive and self-adapted music-listening tests: Psychometric features and motivational benefits. *Applied Measurement in Education, 7*, 25-51.

Vispoel, W.P., Rocklin T.R., & Wang, T. (1994). Individual differences and test administration procedures: a comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7*, 53-79.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Weiss, D.J. (1974). *Strategies of adaptive ability measurement.* Research Report 74-5. Arlington: Personnel and Training Research Programs Office Naval Research.

Weiss, D.J., & Kingsbury, J.J. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Weiss, D.J., & Schleisman, J.L. (1994). Adaptive Testing. In T. Husén & T.N. Postlethwaite (Eds.). *The International Encyclopedia of Education: research and studies* (2nd ed., pp. 48-53). Oxford: Pergamon Press.

Wise, S.L., Plake, B.S., Johnson, P.L., & Roos, L.L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement, 29*, 329-339.

Recent Measurement and Research Department Reports:

97-1 H.H.F.M. Verstralen. A Logistic Latent Class Model for Multiple Choice Items.

97-2 N.D. Verhelst. A New Heuristic for Estimating the Reliability of a Test.

97-3 N.D. Verhelst. Estimating of Latent Abilities from Raw Test Scores.