Computerized Adaptive Testing of Personality Traits

A. Michiel Hol, Harrie C.M. Vorst, and Gideon J. Mellenbergh

University of Amsterdam, The Netherlands

Abstract. A computerized adaptive testing (CAT) procedure was simulated with ordinal polytomous personality data collected using a conventional paper-and-pencil testing format. An adapted Dutch version of the dominance scale of Gough and Heilbrun's Adjective Check List (ACL) was used. This version contained Likert response scales with five categories. Item parameters were estimated using Samejima's graded response model from the responses of 1,925 subjects. The CAT procedure was simulated using the responses of 1,517 other subjects. The value of the required standard error in the stopping rule of the CAT was manipulated. The relationship between CAT latent trait estimates and estimates based on all dominance items was studied. Additionally, the pattern of relationships between the CAT latent trait estimates and the other ACL scales was compared to that between latent trait estimates based on the entire item pool and the other ACL scales. The CAT procedure resulted in latent trait estimates qualitatively equivalent to latent trait estimates based on all items, while a substantial reduction of the number of used items could be realized (at the stopping rule of 0.4 about 33% of the 36 items was used).

Keywords: adaptive testing, cmoputer-assisted testing, item response theory, Likert scales, personality measures

Introduction

In the area of psychological assessment, the development of computerized versions of conventional paper and pencil tests has become increasingly popular (Finger & Ones, 1999; Mead & Drasgow, 1993). The use of computers can improve the efficiency of psychological testing by reducing labor costs, decreasing scoring errors, increasing test standardization, increasing test security, and increasing speed in processing the subjects' responses (Drasgow & Olson Buchanan, 1999). Moreover, computerized psychological assessment introduces the possibility of more advanced test administration procedures, which were impossible to implement in conventional paper-and-pencil testing.

The use of computers in testing has another advantage when item response theory (IRT) is used. In IRT, the trait levels of different subjects can be estimated from different sets of items. Yet, these estimates are comparable to each other. Additionally, it can be determined which item from a set is most informative at a particular trait level. These two characteristics are used in a computerized adaptive test (CAT). In a CAT, each time a person answers an item, his or her trait level is estimated, and the most informative item from the item bank is chosen as the next item to be administered. When the test score precision is sufficient, the computerized administration of items stops. Items that are most informative for a specific subject are administered to this subject.

If a bank of items informative across a wide range of the latent trait scale is created, the latent trait values can be estimated more precisely using a CAT than conventional fixed length paper-and-pencil tests. In a CAT, each time a person is tested, a person-specific subset of items is selected. Therefore, the same amount of questions in a CAT can lead to test scores with a higher precision than in a conventional paper-and-pencil test. Literature shows that CAT exams accomplish measurements with equal or better precision using an average of 50% of the original number of items that is used in paper-and-pencil tests (Embretson & Reise, 2000).

In general, IRT has not found as much application in personality measurement as in educational measurement. But, as indicated in Rouse, Finger, and Butcher (1999), in recent years more studies have appeared in the literature. Nevertheless, current CAT applications typically consist of dichotomously scored cognitive ability items (Dodd, De Ayala, & Koch, 1995). Personality tests consisting of dichotomous items could be implemented in computer programs developed for adaptive testing of cognitive abilities. However, the majority of personality tests consists of polytomous items. IRT models for the analysis of ordinal polytomous data do exist, however, and these can be used for the development of computerized adaptive tests with Likert-type items. Although adaptive testing for personality assessment has been discussed in the literature, the number of such applications is small (Dodd et al., 1995; Reise, 1999; Waller, 1999; Waller & Reise, 1989). The objective of this study is to demonstrate the application of adaptive testing procedures to personality tests having Likert items using Samejima's (1969) graded response model (GRM).

In the GRM, the probability of a response in category x_g of item g as a function of latent trait θ is defined by

Zeitschrift für Psychologie / Journal of Psychology 2008; Vol. 216(1):12–21
DOI 10.1027/0044-3409.216.1.12

$$P_{x_g} \equiv P(X_g = x_g|\theta) = \frac{\exp[a_g(\theta - b_{x_g})]}{1+\exp[a_g(\theta - b_{x_g})]} - \frac{\exp[a_g(\theta - b_{x_g+1})]}{1+\exp[a_g(\theta - b_{x_g+1})]}. \quad (1)$$

Here x_g can take the values 0,. . .,m_g (m_g is the number of category boundaries of item g with m_g+1 categories), a_g is the discrimination parameter of item g ($a_g > 0$) and b_{x_g} is the location parameter of category x_g of item g. Discrimination parameter a_g is constant for all categories of item g but can vary between items. Because all items in this study have five response categories, $m_g = 4$ and $x_g = 0,1,2,3,4$.

For the lowest category $x_g = 0$, it holds that $b_0 = -\infty$ and for the highest category, $x_g = m_g$, it holds that $b_{m_g+1} = +\infty$. Therefore, the probability of a response in the lowest category is given by

$$P(X_g = 0) = P_0 = 1 - \frac{\exp[a_g(\theta - b_1)]}{1+\exp[a_g(\theta - b_1)]}, \quad (2)$$

and the probability of a response in the highest category is given by

$$P(X_g = m_g) = P_{m_g} = \frac{\exp[a_g(\theta - b_{m_g})]}{1+\exp[a_g(\theta - b_{m_g})]}. \quad (3)$$

CAT research can be done using simulation studies, real data simulation studies and studies in real-world testing. In CAT simulation studies, item response data are generated according to a specific model and these data are used to simulate a CAT procedure. In CAT real-data simulation studies, item response data gathered with a conventional fixed length testing method (paper-and-pencil or computerized) are used to simulate a CAT procedure. The responses to the items that would have been selected in a real CAT procedure are used to compute θ estimates. In real-world studies actual subjects respond to a computerized test that administers items adaptively.

Each of the above research strategies has its pros and cons, which however can compensate each other. In the present study, a real-data simulation was done to assess the potential of the application of CAT for personality measurement in real-world applications. Although simulation studies give a researcher more control, the strength of a real-data simulation study is that it can assess whether it is useful to apply theoretical models and their applications to real data. For example, real-data simulation studies can give an indication of the practical relevance of model-data misfit, which can be a problem in psychological assessment. In addition, real-data simulation studies can be done at a lower cost than real-word CAT studies.

A simulation study with the GRM by Dodd, Koch, and Ayala (1989) resulted in recommendations about the stopping rule, the size of the item bank, and the method for estimating an initial θ value. The present study complemented the results and recommendations of this simulation study with a real-data simulation study with polytomous item response data of the dominance scale of the adjective checklist (ACL) (Gough & Heilbrun, 1980). An adapted Dutch version of the ACL dominance scale was used. This version contains items with five ordered polytomous categories (Hendriks, Meiland, Bakker, & Loos, 1995).

Singh, Howell, and Rhoads (1990) also performed a real-data simulation using polytomous item responses. The item bank consisted of 12 items measuring consumer satisfaction. The authors concluded that it was useful to test adaptively because the precision of the θ estimates was sufficiently high after a mean number of eight items. However, according to Dodd et al. (1995) this study was limited by the extremely small item bank of 12 items. The present study extended this work to a larger item bank (36 items) containing items used for personality measurement.

The present study also gives an indication of the validity of adaptive test scores by studying the relation between the adaptive test scores and the test scores of the other ACL scales. In the adaptive test the required precision of the final θ estimates was manipulated. The effect on the mean number of items required, the correlation of the adaptive test scores with conventional test scores as well as the other ACL scales was studied. In this way, the efficiency of adaptive testing of polytomous personality scales could be assessed. A similar study using paper-and-pencil data with a Dutch personality scale measuring neuroticism was conducted by Hol, Vorst, and Mellenbergh (2001); this study showed that it was possible to obtain efficient adaptive personality trait estimates at a substantial reduction of the number of items. Also, the correlation between the adaptive trait estimates and trait estimates based on the entire item pool remained reasonably high. In addition, the pattern of correlations between adaptive trait estimates and other personality scales remained very similar to the pattern of correlations for trait estimates based on the entire item pool (Hol et al., 2001).

Method

Participants

All psychology freshmen at the University of Amsterdam must take a number of paper-and-pencil tests as an obligatory part of their study commitments. The data for the dominance scale of the ACL from 3,587 participants of the years 1993 through 2000 were used (29.8% men and 69.2% women, mean age = 21.6 years). The data from the years 1993–1997 (N = 1995) were used to estimate the item parameters, which must be known in a real-data simulation of a CAT procedure. The data from 1998–2000 were used for the real data simulation of the CAT procedure (N = 1592). The order of item calibration and simulated operational CAT is similar to that in real-world applications.

Materials

The ACL (Gough & Heilbrun, 1980) is a personality questionnaire most commonly used to obtain self-descriptions.

It can be used both for general and clinical populations. The items in the ACL are adjectives for which the person is instructed to judge the applicability of them to himself or herself.

The common version of the ACL consists of dichotomous items, but in this study an adapted Dutch version of the ACL was used which instructs the persons to judge the applicability of the adjectives to themselves on a 5-point Likert scale (Hendriks et al., 1995).

In the real-data simulation, the ACL dominance scale consisted of 40 items. However, four of these items had to be eliminated (see the Results section). A typical example of an ACL dominance item is:

Strong: 0–1–2–3–4

The ACL items are listed together on multiple pages. For each adjective, the person judges the applicability and puts a circle around the appropriate number. The meanings that belong to these numbers, which range from "Not applicable at all" (0) to "Fully applicable" (4), are repeated on the top of each page.

Item Parameters and Unidimensionality Assumption

The administration of a CAT requires the item parameters to be known. Therefore, they were estimated beforehand, using *Multilog* (Thissen, 1991) and assuming a normally distributed latent trait with mean 0 and standard deviation 1.

Two versions of the GRM were tested against each other using the data from 1993–1997 ($N = 1995$). The GRM with equal discrimination parameters (a) for all items was tested against the GRM that allows the parameters to differ across items (Bock & Lieberman, 1970, p. 194). The Likelihood Ratio (LR) test showed that the a-parameters could not be set equal to each other ($LR(35) = 1293.9$, $p < .05$).

For acceptable item calibration, Reckase (1979) recommended a dominant first factor that accounts for at least 20% of the test variance. An exploratory factor analysis with the program *Microfact* (Waller, 2003) on the polychoric correlations between the item variables showed a first dominant factor that accounted for 24% of the test variance.

Adaptive Procedures

The latent trait values in the adaptive procedure were estimated using the maximum-likelihood method. The standard errors of the maximum-likelihood estimates are estimated by taking the root of the negative inverse of the second derivative of the loglikelihood function as is done in *Multilog* (Thissen, 1991, sec. 4–100). These errors are estimated as a by-product of maximum-likelihood estimation.

It is not possible to determine maximum-likelihood estimates for response patterns having identical extreme responses to each of the test items (e.g, patterns of responses in the first category for all items, or patterns of responses all in the last category). Therefore, the data of persons with such patterns were deleted from the 1998–2000 data. Although no maximum-likelihood θ estimates can be calculated for these persons, their results are not meaningless in real-world testing. These persons just have θ values that probably exceed the smallest or largest b-parameter of the scale.

A computer program was written to perform a real data simulation of a CAT with ordinal polytomous item response data. The program used the logistic version of the GRM (Samejima, 1969, 1997). The item information functions (Samejima, 1969, p. 39) were used to select the items in the adaptive procedure. The item information function was derived by Samejima to be

$$I_g(\theta) = \sum_{x_g=0}^{m_g} \frac{\left(P'_{x_g}\right)^2}{P_{x_g}}, \tag{4}$$

where P_{x_g} is the probability of a response in category x_g of item g ($x_g = 0, 1, 2, 3, 4$) as given by Formulas 1, 2, and 3, respectively. P'_{x_g} is the first derivative of P_{x_g} with respect to θ. Every time a new item has to be selected, the item with the largest information value at the current θ estimate is picked from the free items in the pool.

At the beginning of a CAT, no empirical θ estimate is available. Therefore, an initial θ value has to be selected. Without a priori information, when a symmetrical normal ability distribution with a mean equal to 0 can be assumed, the best initial value for an arbitrary selected subject is 0 This choice was also made in the current study. Hence, the item had the largest item information value at θ = 0. This item was thus the same for each of the subjects.

If the first response of the subject is on the extreme negative or positive side of the scale, the trait estimate would take the value of –∞ or ∞, respectively. Then, it is impossible to select the second item using the item information functions. Therefore, as proposed by Dodd et al. (1989), the initial value should be recalculated using a variable stepsize method. In the present study, a slightly different version of the Dodd et al. (1989) method was used, which computes the new initial value as the mean of the earlier initial value (0) and the largest (lowest) b-parameter when the response is on the extreme positive (negative) side of the scale. The second item is then selected at this new value. When a subject continued to respond in the same extreme response category, again a new initial value was computed from the last initial value and the smallest or largest location parameter. This procedure of computing initial values was continued until the subject leaves the extreme category. At that point a maximum-likelihood estimate of θ was computed and the next item was selected to be optimal at this estimate.

If the first response was not extreme, the CAT just calculated the θ estimate and the standard error of it. This new θ estimate was used to compute the value of the item infor-

mation function of all remaining items. The item with the largest value of it at the estimate was selected as the second item for the subject, and so forth.

The procedure continued until the latent trait was estimated with sufficient precision according to its estimated standard error. In the computer program, a maximum value of the standard error of the latent trait was set. The CAT stopped administering items to a subject when the standard error of the θ estimate dropped below this maximum value.

Manipulation of the Stopping Rule

The maximum standard error defined in the stopping rule was manipulated in the simulations. The adaptive procedure was run six times with the following settings of the maximum standard error: 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8.

Analysis

The effect of the manipulation of the stopping rule was studied in three ways. First, the correlations of the adaptive trait estimates with the trait estimates from the responses to all 36 items were examined. Second, the correlations between the adaptive trait estimates and 33 other ACL scales were compared with those between estimates based on all 36 items and the 33 other ACL scales. Third, the number of items that was used for the different stopping rule settings was investigated.

Results

Parameter Estimation

Data of 1995 persons were available for item parameter estimation. 70 of them had more than 10% of their responses on the ACL items missing. These persons were ignored, and the item parameters were thus estimated from the responses of 1925 persons. Originally, the dominance scale consisted of 40 items. Four items could not be used because these items had very small *a*-parameters. These items would not have added much information to the θ estimates. The estimates of the item parameters of remaining 36 items are shown in Table 1.

Adaptive Procedures

For the adaptive procedure, the data of 1592 persons were available. The program that was written to mimic the adaptive procedure requires complete answer patterns. Fifty-five persons had more than 10% missing responses on the ACL items; 20 persons had answer patterns with all responses in one extreme category. These persons were eliminated from the data set.

Figure 1 shows the relationship between the θ estimates (n = 1517) based on the responses on all 36 items as well as their estimated standard errors. The standard errors are larger at the right side of the scale, which was probably caused by the distribution of the *b*-parameters.

Table 1 shows that the lowest location parameters (*b1*) are more extreme at the negative side of the scale than the highest location (*b4*) parameters at the positive side of the scale. This causes larger positive θ estimates to have larger standard errors.

The first item used in the adaptive procedure was Item 36. This item was chosen because it had the largest value of the item-information-function at θ = 0. The value was caused by a large *a*-parameter and a location parameter – b3 – very close to 0 for this item (see Table 1). Recent literature on CATs for the measurement of abilities shows that administering tests from the same item bank can cause high exposure rates for certain items. In general, items with large *a*-parameters are exposed more often than items with small *a*-parameters (Chang, Qian, & Ying, 2001). Figure 2 shows the relationship between *a*-parameters and their usage percentage for the stopping rule with standard error of estimation set at 0.4. There is an apparent relationship between the percentage of item use and the *a*-parameters: Items with large *a*-parameters are used more often than items with small *a*-parameters.

Table 2 shows for each stopping rule (i) the mean number of items used, (ii) the correlation between the adaptive θ estimates and the estimates based on the entire set of 36 items, and (iii) the mean standard errors of the adaptive θ estimates. When the standard error in the stopping rule increases, the mean number of used items decreases, the correlation of adaptive θ estimates with θ estimates based on the entire set of items decreases but the mean of the standard errors of the adaptive θ estimates increases. The mean of the standard errors was 0.306 for the stopping rule with errors smaller than 0.3 (the mean was larger than the threshold because some of the persons who reached the last item in the pool still have standard errors greater than 0.3). The use of all 36 items only occurred for the stopping rules with the smallest standard errors (0.3 and 0.4). For the former, 456 persons (30.1%) used all items; for the latter, only 6 persons (0.4%).

The correlations between adaptive θ estimates and estimates based on the entire set of items remained high over the range of stopping rules studied. However, this result was as expected. First, adaptive θ estimates are based on the same data as the estimates from the entire set of items. Second, θ estimates with large standard errors caused the adaptive procedure to use all items for these estimates; therefore, these estimates were the same as those directly based on the entire set of items. θ estimates with large standard errors are often extremely positive or extremely negative. Therefore, the correlations for the persons in the CAT procedure with θ estimates

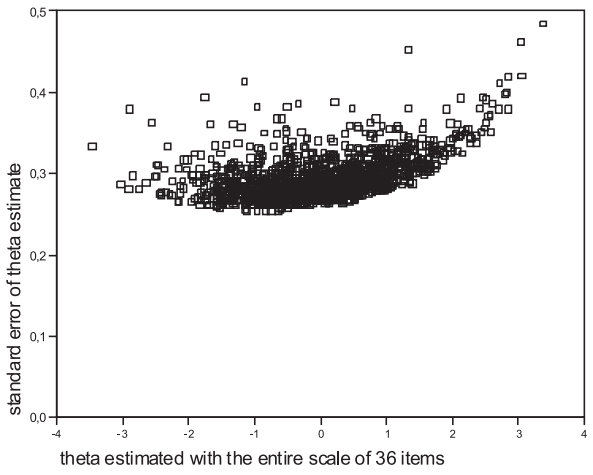

Figure 1. Relationship between trait estimates determined with the entire scale and the standard errors of these trait estimates.

Table 1. Item a-parameters and b-parameters (b1, b2, b3, b4)

Item	a	b1	b2	b3	b4
1	1.10	–4.22	–2.08	0.04	2.12
	(0.07)	(0.33)	(0.14)	(0.06)	(0.14)
2	0.89	–5.59	–2.81	–0.21	3.25
	(0.06)	(0.49)	(0.20)	(0.07)	(0.23)
3	0.83	–5.35	–2.81	–0.45	2.36
	(0.06)	(0.46)	(0.21)	(0.08)	(0.17)
4	1.09	–5.96	–4.03	–1.73	–0.14
	(0.07)	(0.66)	(0.33)	(0.13)	(0.06)
5	0.80	–4.18	–1.91	0.76	3.49
	(0.06)	(0.36)	(0.16)	(0.10)	(0.24)
6	1.55	–2.82	–1.36	0.25	1.96
	(0.08)	(0.15)	(0.08)	(0.05)	(0.10)
7	0.39	–8.35	–3.03	0.40	5.59
	(0.06)	(1.26)	(0.40)	(0.17)	(0.73)
8	0.73	–5.81	–2.53	–0.32	2.34
	(0.06)	(0.55)	(0.24)	(0.09)	(0.21)
9	1.11	–4.79	–2.53	–0.25	2.37
	(0.07)	(0.39)	(0.15)	(0.06)	(0.14)
10	0.57	–3.06	–0.27	2.11	6.11
	(0.06)	(0.30)	(0.12)	(0.19)	(0.57)
11	0.53	–3.82	–0.23	2.28	5.48
	(0.06)	(0.45)	(0.12)	(0.26)	(0.55)
12	1.26	–4.49	–2.21	–0.18	2.53
	(0.07)	(0.39)	(0.15)	(0.09)	(0.25)
13	1.10	–5.78	–2.78	–1.44	0.62
	(0.07)	(0.61)	(0.18)	(0.10)	(0.07)
14	1.09	–4.21	–2.03	0.15	2.88
	(0.07)	(0.32)	(0.13)	(0.06)	(0.17)
15	1.50	–3.54	–1.51	0.00	1.87
	(0.07)	(0.23)	(0.08)	(0.05)	(0.10)
16	0.78	–6.70	–3.95	–1.92	0.70
	(0.06)	(0.76)	(1.13)	(0.18)	(0.11)
17	0.69	–4.18	–1.49	0.49	2.64
	(0.06)	(0.39)	(0.16)	(0.10	(0.24)
18	0.43	–9.26	–4.87	–1.22	4.11
	(0.07)	(1.33)	(0.64)	(0.18)	(0.43)
19	1.09	–3.03	–0.89	0.62	2.46
	(0.07)	(0.19)	(0.08)	(0.07)	(0.15)
20	0.96	–4.42	–2.29	–0.43	2.31
	(0.07)	(0.33)	(0.16)	(0.07)	(0.17)
21	1.10	–4.48	–1.91	0.03	2.21
	(0.07)	(0.38)	(0.13)	(0.06)	(0.14)
22	0.62	–6.37	–2.59	0.48	4.41
	(0.06)	(0.62)	(0.24)	(0.11)	(0.37)
23	0.87	–5.62	–3.48	–1.31	1.90
	(0.06)	(0.50)	(0.26)	(0.11)	(0.16)
24	1.48	–3.28	–1.39	–0.01	1.74
	(0.08)	(0.20)	(0.08)	(0.05)	(0.09)
25	1.41	–3.31	–1.74	–0.07	2.06
	(0.07)	(0.20)	(0.09)	(0.05)	(0.11)
26	0.94	–4.77	–2.31	–0.59	1.50
	(0.06)	(0.40)	(0.17)	(0.08)	(0.12)
27	1.43	–2.61	–0.98	0.32	1.84
	(0.07)	(0.28)	(0.13)	(0.11)	(0.21)
28	1.40	–3.39	–1.62	–0.29	1.19
	(0.07)	(0.21)	(0.09)	(0.05)	(0.08)
29	1.38	–4.64	–3.07	–1.77	–0.16
	(0.08)	(0.43)	(0.19)	(0.11)	(0.05)
30	1.25	–4.04	–2.16	–0.25	2.25
	(0.07)	(0.31)	(0.13)	(0.06)	(0.13)
31	1.28	–4.87	–2.91	–1.48	0.22
	(0.08)	(0.45)	(0.18)	(0.09)	(0.06)
32	1.03	–4.76	–2.25	–0.41	1.70
	(0.06)	(1.19)	(****)[a]	(****)	(0.12)
33	0.90	–5.07	–2.63	–1.16	0.68
	(0.06)	(0.45)	(0.21)	(0.11)	(0.09)
34	0.57	–8.24	–4.94	–1.43	1.87
	(0.25)	(1.13)	(0.57)	(0.20)	(0.22)
35	1.67	–4.14	–2.71	–1.34	0.42
	(0.09)	(0.36)	(0.14)	(0.07)	(0.05)
36	1.64	–2.89	–1.28	–0.10	1.48
	(0.08)	(1.36)	(0.12)	(0.08)	(0.14)

Note. Standard errors of the item parameters are between parentheses.
[a](****): Multilog (Thissen, 1991) could not determine standard errors of these item parameters.

Table 2. Effect of the manipulation of the stopping rule

Stopping rule	Mean amount of items used	Correlation adaptive θ estimates with θ estimates based on the entire scale	Mean standard errors
No stopping rule	36	1	0.293
SE(θ) < 0.3	28.13	0.996 (0.994)[a]	0.306
SE(θ) < 0.4	11.74	0.949 (0.932)	0.393
SE(θ) < 0.5	6.83	0.895 (0.868)	0.480
SE(θ) < 0.6	4.79	0.881 (0.852)	0.562
SE(θ) < 0.7	3.60	0.869 (0.837)	0.641
SE(θ) < 0.8	2.71	0.827 (0.792)	0.727

[a]The values between parentheses are correlations of adaptive θ estimates with θ estimates based on the entire test when the data was restricted to θ estimates in the range of –2 and +2 (N = 1449).

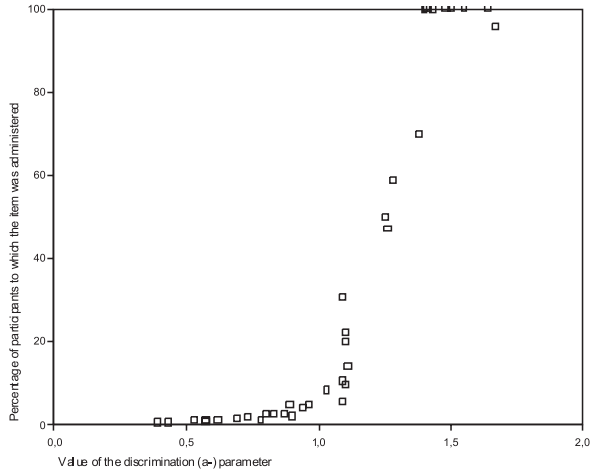

Figure 2. Relationship between the value of the a-parameters of the items and the percentage of subjects to which the item is administered for stopping rule setting 0.4.

from –2 to +2 (N = 1449) on the entire set of 36 items were computed separately. In Table 2, the correlations are shown in parentheses next to the original correlations. Regardless of this selection, the correlations remained about the same for all stopping rules.

The relationship between the θ estimates based on all items and the adaptive estimates for the six different stopping rules are shown in Figures 3(a), 3(c), . . ., 3(k). As these figures indicate, more variation appears when the precision of adaptive trait estimates becomes smaller. That is, when the standard error becomes higher, the relationship between the θ estimates based on all items with the adaptive estimates becomes weaker. Likewise, for each stopping rule, Figures 3(b), 3(d), . . ., 3(l) show the relationships between the adaptive θ estimates and the numbers of items that were used. Figure 3(b) shows that the number of used items is maximal for the extreme trait estimates at the lower side of the scale. In the middle of the scale, the entire item bank was also used for several persons but for the majority of them the number of items was smaller. The adaptive θ estimates that show a deviation from the estimates based on the entire scale are those obtained at a reduced number of items.

In general, when the CAT uses more items, the correlation between the adaptive θ estimates and those based on the entire item set gets higher. Therefore, the scatter plot in Figure 3(a) shows least variation while the plot in Figure 3(k) shows most variation. The difference between the two types of θ estimates is lower at an extreme side of the scale (particularly the right-hand side) than in the middle of it. Extreme θ estimates have larger standard errors; hence the CAT stopped after administration of relatively many items and in the stopping rules 0.3 and 0.4 the CAT sometimes used the entire item bank.

The correlations between both the adaptive θ estimates and those based on the entire item set and the other scales of the ACL are shown in Table 3. The ACL originally had some items that were included in multiple scales. To prevent spurious correlations, in this study the items that figure in the ACL dominance scale were eliminated from all other scales.

Table 3 shows that the pattern of correlations between the adaptive θ estimates and the sum scores of the ACL personality scales remained the same. In general, the absolute correlations became smaller when the standard error in the stopping rule increased. The last column of Table 3 shows the mean of all absolute correlations after Fisher-z transformation. The stopping rules that require standard errors to be smaller than 0.3 and 0.4 show correlations that do not differ very much from the original correlations. Note that the mean number of items used in stopping rule with the standard error below 0.3 was 28.13; the mean number of items in stopping rule with standard errors below 0.4 was only 11.74.

Because of their lower reliability, the CATs with higher standard error settings in Table 2 and Table 3 will have "true" correlations (i.e., after correction for attenuation) that are higher than the reported values.

Table 3. Correlations of the dominance scale with other scales of the ACL under different stopping rule settings

	Sum scores of ACL personality scales																
Stopping rule	Ach	End	Ord	Int	Nur	Aff	Het	Exh	Aut	Agg	Cha	Suc	Aba	Def	CrsM	CrsF	S-Cn
All items	.544	.388	.315	.479	.414	.612	.633	.213	–.074	–.289	.201	–.340	–.562	–.151	–.804	–.119	–.188
SE(θ) < 0.3	.518	.367	.291	.469	.419	.608	.637	.215	–.081	–.295	.205	–.337	–.558	–.150	–.803	–.136	–.191
SE(θ) < 0.4	.401	.273	.196	.375	.381	.538	.616	.249	–.074	–.259	.186	–.307	–.550	–.161	–.755	–.161	–.228
SE(θ) < 0.5	.333	.218	.143	.297	.339	.460	.571	.251	–.078	–.232	.153	–.273	–.517	–.154	–.696	–.181	–.224
SE(θ) < 0.6	.344	.241	.162	.318	.347	.474	.580	.225	–.086	–.244	.143	–.275	–.506	–.138	–.692	–.180	–.198
SE(θ) < 0.7	.362	.266	.190	.339	.346	.493	.572	.215	–.074	–.254	.123	–.304	–.516	–.139	–.688	–.158	–.178
SE(θ) < 0.8	.338	.258	.184	.314	.351	.451	.533	.164	–.112	–.257	.111	–.303	–.486	–.091	–.643	–.192	–.146

	Sum scores of ACL personality scales															
Stopping rule	S-Cfd	P-Adj	Iss	Cos	Mls	Mas	Fem	CP	NP	A	FC	AC	A-1	A-2	A-3	A-4	M
All items	.741	.588	.687	.506	.497	.142	.327	–.127	.522	.549	.741	–.587	.249	–.096	.422	.292	.431
SE(θ) < 0.3	.741	.587	.682	.500	.486	.129	.320	–.143	.522	.538	.748	–.579	.244	–.102	.418	.272	.428
SE(θ) < 0.4	.686	.529	.621	.461	.396	.087	.260	–.164	.474	.461	.724	–.508	.208	–.125	.363	.179	.381
SE(θ) < 0.5	.621	.465	.569	.413	.329	.044	.194	–.180	.429	.410	.682	–.450	.166	–.156	.299	.105	.336
SE(θ) < 0.6	.621	.478	.579	.417	.347	.042	.206	–.185	.440	.421	.671	–.463	.172	–.152	.310	.120	.341
SE(θ) < 0.7	.623	.485	.600	.426	.362	.074	.216	–.181	.452	.443	.664	–.489	.196	–.136	.322	.160	.349
SE(θ) < 0.8	.592	.467	.560	.371	.370	.024	.193	–.197	.443	.426	.613	–467	.143	–.184	.283	.110	.326

Note. Ach = achievement; End = endurance; Ord = order; Int = intraception; Nur = nurturance; Aff = affiliation; Het = heterosexuality; Exh = exhibition; Aut = autonomy; Agg = aggression; Cha = change; Suc = succorance; Aba = abasement; Def = deference; Crs-M = counseling readiness males; Crs-F = counseling readiness females; S-Cn = self-control; S-Cfd = self-confidence; P-Adj = personal adjustment; Iss = ideal self; Cps = creative personality; Mls = military leader; Mas = masculine attributes; Fem = feminine attributes; CP = critical parent; NP = nurturant parent; A = adult; FC = free child; AC = adapted child; A-1 = hi origence lo intellectance; A-2 = hi origence hi intellectance; A-3 = lo origence lo intellectance; A-4 = lo origence hi intellectance; M = mean of all absolute correlation values after Fisher-*Z* transformation.

Discussion

As reconfirmed by this research, adaptive θ estimates with a sufficient precision can be obtained from numbers of items that were substantially smaller than the total number of items in the entire scale. The stopping rule that required the standard errors to be smaller than 0.3 needed an average of 78% of the original number of items. Nevertheless, the correlation between the adaptive estimates and the estimates based on the entire set of items remained very high (.99). The stopping rule that required the standard errors to be smaller than 0.4 needed an average of only 33% of the entire set of items while the same correlation still was .95. This last rule was the best in this study. It maintained a very large correlation between the two types of estimates while the number of items needed was reduced substantially.

In an earlier study with a scale measuring neuroticism it was found that a stopping rule that required the standard errors to be smaller than 0.5 performed best (Hol et al., 2001).

In the current study, we expected the correlations between the adaptive θ estimates and the estimates based on the set of items to be high because both types of estimates were based on the same data. Therefore, we also correlated the adaptive θ estimates with the scores on the other scales in the ACL. Items that also figure in the dominance scale were deleted from these other scales. The correlations between the adaptive θ estimates and the scores on the other scales stayed close to those based on the entire sets of items for each scale. This was especially true for the stopping rules that required standard errors to be smaller than 0.3 or 0.4. In general, as expected, the absolute values of the correlations decreased when the allowed standard error of the stopping rule increased. The fact that the adaptive estimates were from different subsets of items did not have any significant impact on the correlations.

When adaptive testing is applied in real-world testing, it might lead to some loss of predictive value but at a substantial reduction of the number of items needed. Additionally, this loss is not necessary. In general, when a personality test is well designed for a population of test takers, it is informative at θ values close to the mean of the population. For regular populations, paper-and-pencil tests can therefore be informative for a large proportion of it. But for such a test to be informative for a wider range of the trait too many items would be required. In this type of applications, one could resort to adaptive testing provided, of course, an item-bank can be created that systematically covers the entire range of the personality trait that is measured. Future research should be aimed at how to create such item banks. The same conclusion was made by Koch, Dodd, and Fitzpatrick (1990) in a study of computerized adaptive administration of an attitude scale.

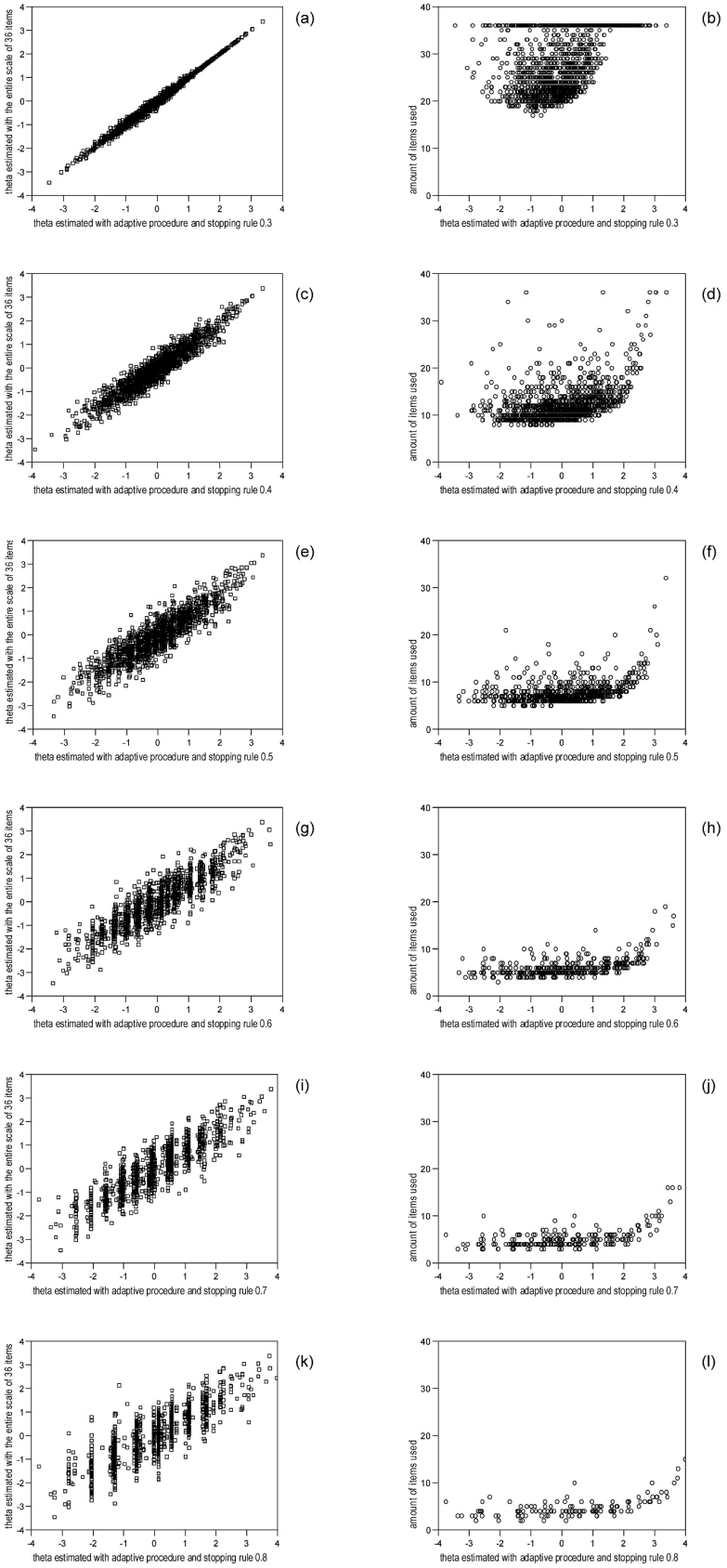

Figure 3. The effect of the manipulation of the stopping rule on the quality of the adaptive trait estimates and the number of items that is used.

Research on cognitive ability tests shows that adaptive tests, computerized conventional tests, and paper-and-pencil tests lead to equivalent scores. A meta-analysis conducted by Mead and Drasgow (1993) supported the conclusion for paper-and-pencil tests compared to computerized tests. Furthermore, no significant differences were found between computerized conventional and computerized adaptive tests. It should be noted, however, that an effect of administration mode was found for the degree of speediness of ability tests. But speed is usually not an important factor in personality questionnaires.

In another meta-analysis (Finger & Ones, 1999), the equivalence of computerized and booklet forms of the MMPI was studied. Finger and Ones (1999) concluded that both were equivalent indeed, and that norms estimated for booklet forms can be safely used with computerized versions. Future research should address the question of the equivalence of IRT based computerized adaptive and paper-and-pencil personality tests.

In some tests, items that are positively formulated with respect to the trait are balanced with items that are negatively formulated. This strategy is used to avoid the impact of response styles of participants (Oosterveld & Vorst, 1998). In a CAT, some θ estimates may be based on either positive or negative items. The probability of this happening may become higher for a shorter test. In principle, such events should not be any serious if the IRT model fits. But if one wants to avoid them, a form of content balancing should be build into the adaptive item selection algorithm.

The same items can be administered in a different order in an adaptive test (Singh et al., 1990). Traditionally, it has been hard to study possible order effects because the items in paper-and-pencil tests are administered in a fixed order. But this problem does not exist for computerized adaptive testing.

Furthermore, in adaptive testing with a fixed standard error of measurement, the number of items administered to each person is no longer fixed. Hence, for longer adaptive tests, possible effects of fatigue and boredom can be different across persons. Such effects should also be studied.

The literature about adaptive testing of abilities shows that there is a risk of higher exposure rates for items with large *a*-parameters because of the positive influence of this parameter on the information function. This effect did also occur in this study. High exposure rates of ability items can lead to familiarity with these items in the population of examinees when the tests are applied continuously (*walk-in testing*). In recent studies, methods have been developed to avoid high exposure rates of specific items by restricting item selection procedures in CATs (Chang et al., 2001; Meijer & Nering, 1999; van der Linden & Reese, 1998). However, this problem is less relevant to the measurement of personality traits than of abilities.

In conclusion, this study indicates that there is potential for the development of adaptive tests for personality assessment. Although the number of studies in this domain increases slowly but steadily, much work is still to be done.

Acknowledgments

The first author currently works at the Dutch Probation Service but this research was done as part of his dissertation project at the University of Amsterdam. This project was supported by a grant from the Netherlands Organization for Scientific Research (NWO).

References

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35,* 179–197.

Chang, H., Qian, J., & Zhiliang, Y. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25,* 333–341.

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

Dodd, B.G., Koch, W.R., & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13,* 129–143.

Drasgow, F., & Olson Buchanan, J.B. (Eds.). (1999). *Innovations in computerized assessment.* Mahwah, NJ: Erlbaum.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Finger, M.S., & Ones, D.S. (1999). Psychometric equivalence of the computer and booklet forms of the MMPI: A meta-analysis. *Psychological Assessment, 11*, 58–66.

Gough, H.G., & Heilbrun, A.B. (1980). *The Adjective Check List manual.* Palo Alto, CA: Consulting Psychologists Press.

Hendriks, C., Meiland, F., Bakker, M., & Loos, I. (1995). *Eenzaamheid en persoonlijkheidskenmerken* [Loneliness and personality attributes]. Unpublished manuscript, University of Amsterdam, The Netherlands.

Hol, A.M., Vorst, H.C.M., & Mellenbergh, G.J. (2001). Toepassing van een computergestuurde adaptieve testprocedure op persoonlijkheidsdata [Application of a computerized adaptive test procedure on personality data]. *Nederlands Tijdschrift voor de Psychologie, 56,* 119–133.

Koch, W.R., Dodd, B.G., & Fitzpatrick, S.J. (1990). Computerized adaptive measurements of attitudes. *Measurement and Evaluation in Counseling and Development, 23*, 20–30.

Mead, A.D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449–458.

Meijer, R.R., & Nering, M.L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23,* 187–194.

Oosterveld, P., & Vorst, H.C.M. (1998). Constructie van meetinstrumenten [Construction of measurement instruments]. In W.P. van den Brink, & G.J. Mellenbergh (Eds.), *Testleer en testconstructie* [Test theory and test construction] (pp. 303–337). Amsterdam: Boom.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207–230.

Reise, S.P. (1999). Personality measurement issues viewed through the eyes of IRT. In S.E. Embretson, & S.L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219–241). Mahwah, NJ: Erlbaum.

Rouse, S.V., Finger, M.S., & Butcher, J.N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72*, 282–307.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Samejima, F. (1997). Graded response model. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Singh, J., Howell, R.D., & Rhoads, G.K. (1990). Adaptive designs for Likert-type data: An approach for implementing marketing research. *Journal of Marketing Research, 27,* 304–321.

Thissen, D. (1991). *Multilog 6.0 user's guide.* Chicago: Scientific Software.

Van der Linden, W.J., & Reese, L.M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22,* 259–270.

Waller, N.G. (1999). Searching for structure in the MMPI. In S.E. Embretson, & S.L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know.* (pp. 185–217). Mahwah, NJ: Erlbaum.

Waller, N.G. (2003). *MicroFACT: A microcomputer factor analysis program for ordered polytomous data and mainframe size problems* (Version 2.1) [Computer Software]. St. Paul, MN: Assessment Systems Corporation.

Waller, N.G., & Reise, S.P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption Scale. *Journal of Personality and Social Psychology, 57,* 1051–1058.

A. Michiel Hol

The Netherlands
Tel. +31 20 400-4077
Fax +31 842269244
E-mail a.michiel.hol@adaptief.nl

