Graduate
Management
Admission
Council®

The Development of a Computerized Adaptive Test for Integrity

Iris J. L. Egberink

Bernard P. Veldkamp

University of Twente

Presented at the June 8, 2007 Poster Session

Abstract

Integrity tests are used widely in personnel selection because they have incremental validity to cognitive ability tests (e.g., Ones & Viswesvaran, 2002) and they have been shown to predict organizational outcomes such as job performance and absenteeism (e.g., Ones, Viswesvaran, & Schmidt, 1993). In practice, most integrity tests are administered by paper-and-pencil (P&P). However, administration by means of a computer is becoming more popular because of the efficient way tests can be administered and the fact that they can be administered online (e.g., Naglieri et al., 2004). Using a computer to administer items gives the possibility for adaptive administration. The aim of the present study was to develop a computerized adaptive version of an integrity test that can be administered online and to investigate whether this adaptive version resulted in more efficient measurement than the classical P&P approach. Because it takes more time and money to develop and maintain a CAT as compared to P&P tests, it is important to investigate the efficiency of a CAT relative to linear tests. First results showed that reliable measurement, was possible at lower levels of integrity with reduction of 50% of the items. However, only a small part of the item pool was used. This illustrates one of the inherent problems of personality-based CATs:a few good items tend to dominate the test.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2007 by the Authors

Citation

Egberink, I. J. L. & Veldkamp, B. P. (2007). The development of a computerized adaptive test for integrity. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Iris J. L. Egberink, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: i.j.l.egberink@utwente.nl

The Development of a Computerized Adaptive Test for Integrity

Integrity tests are widely used in personnel selection. These tests are used for assessing trustworthiness, honesty, dependability, reliability, and conscientiousness of employees and job applicants. The aim of these tests is to predict a variety of counterproductive work behaviors, such as theft and absenteeism (e.g. Ones, 1993; Ones & Viswesvaran, 1998, 2001; Wanek, 1999) One reason for their wide use is their incremental validity to cognitive ability tests (e.g., Ones & Viswesvaran, 2001). Another reason is that integrity tests have shown to predict organizational outcomes such as job performance and absenteeism (e.g., Ones, Viswesvaran, & Schmidt, 1993).

In practice, integrity—or actually the lack of integrity—can have quite an impact on organizations. For example, when a security guard steals from the company which he/she is supposed to secure, the security company is in trouble. In addition to the fact that a trial might cost much money and time, the company needs to handle the negative publicity and they have to find a new employee.

These are the reasons that a large Dutch security company was interested in developing a new instrument for the recruitment and selection of employees. They wanted to have a test that assesses integrity in detail, but in such a manner that it would not be obvious that the test is about measuring integrity. They also wanted to have a test that is not time consuming, because of the other parts of the recruitment and selection process. They defined integrity as the degree to which we can trust that someone does what he/she had to do, says what he/she thinks and does what he/she agrees. The current study was implemented within this context.

In general, there are two types of integrity tests: overt and personality-based integrity tests (Sackett, Burris, & Callahan, 1989). Overt integrity tests (or “clear-purpose” tests), assess directly attitudes regarding theft and prior dishonest behavior. An example is the Reid Report (Reid, 1967). Personality-based integrity tests (or “disguised-purpose” tests) aim to predict a variety of counterproductive behaviors at work based on personality attributes, such as conscientiousness and dependability. An example of this type of integrity tests is the Employee Reliability Inventory (e.g. Sackett, et. al., 1989; Cullen, & Sacket, 2003; Ones, et al., 1993). We choose to develop a personality-based integrity test.

In practice, most integrity tests are administered by paper-and-pencil. However, administration by means of a computer is becoming more popular because of the efficient way questionnaires can be administered and the fact that they can be administered online (e.g., Naglieri et al, 2004). Using a computer gives the possibility of administering items by means of a computerized adaptive test (CAT), that is, choose the items that most adequately measure the trait level of an individual candidate. An advantage is that shorter tests can be administered that measure with similar measurement precision than paper-and-pencil tests or computer-based tests that are administered by computer without using an adaptive selection algorithm for the selection of items (e.g., Embretson & Reise, 2000). This is why we advised that the test should be developed as a CAT.

The use of CAT has become popular in the context of educational measurement, but recently there have appeared applications in the personality domain as well (e.g., Reise & Henson, 2000; Fliege et al.2005; Simms, & Clark, 2005). However, although a shorter test is an advantage of CAT (e.g., Embretson & Reise, 2000; Waller & Reise, 1989), some authors suggest that instead of using a CAT, it could be useful to assemble a shorter fixed-length test that consists of items that provide the most psychometric information and have the highest measurement precision (Reise & Henson, 2000).

This is why we advised to administer the test adaptively. Research was needed to find out whether it would me more efficient to use a CAT. The aim of the present study was to develop a CAT version of a personality-based integrity test, which can be administered online, and can be applied for personnel selection. Besides, it was assessed whether adaptive administration of the test was an efficient to do so. The first study is a description of the construction of the CAT for integrity. The second study describes the results of applying the new instrument.

STUDY 1

Method

Participants

We used a sample of 984 Dutch applicants, who were administered the Workplace Big Five questionnaire (WB5; Schakel, Smid, & Jaganjac, 2007) as part of a selection and recruitment procedure. The WB5 is a personality questionnaire applied to situations and behavior in the workplace. It consists of 144 items, distributed over five scales (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness). The items are scored on a five-point Likert scale. The answer most indicative for the trait being measured is scored 5 and the answer least indicative for the trait is scored 1. The questionnaire was administered online. Coefficient alpha for the whole questionnaire was .78. Coefficient alpha varied from .79 to .91 for the five scales.

The persons in this sample had a mean age of 39.6 (SD = 9.7). There were 57.7% mostly White men and 42.1% mostly White women (for 2% of the persons gender was unknown). 20.8% of them had a university education, 43.8% had higher vocational education and 18.5% secondary vocational education.

Development of the CAT

Construction of the item bank. Two content experts defined five aspects of integrity: the degree to which someone informs others carefully, takes responsibility, respects others, takes control and shows loyalty. For each aspect they took a close look at the 144 items measuring the Big Five, to select items for measuring that aspect. Ultimately they selected 74 (unique) items with overlap in the aspects.

Item responses on these 74 items were input into a confirmatory factor analysis. We conducted a five-factor, four-factor and three-factor analysis. The results of the three-factor analysis were most satisfying. With these results we went back to the content experts. They took a close look again and decided that they agreed with this structure with regard to the content of the factors, and that the 3-factor structure was a good solution. Factor 1 was labeled “Cognitive Stress:” it assesses the degree to which someone can handle unexpected situations and complex problems. The second factor was labeled “Social Stress:” it assesses the degree to which someone speaks without hesitation and coordinates with others. Factor 3 was labeled “Achievement Orientation:” it assesses the degree to which we are purposeful for achieving results.

In order to extend the three scales even further, an empirical approach was applied to determine whether it would be possible to add other items from the WB5 to the integrity scales.

Based on the initial set, we used the search option in the Mokken Scale Analysis for Polytomous Items program (MSP 5.0; Molenaar & Sijtsma, 2000) to select unused items that covaried positively with those items and together formed a scale with $H \geq .25$ (e.g., Sijtsma & Molenaar, 2002). Although $H \geq .3$ is often recommended as a lower bound, too many items were rejected using this lower bound. Therefore, $H = .25$ was chosen. In close cooperation with the content experts, it was decided whether the additional items would be added to the item bank.

To select items that together formed a unidimensional scale, we checked the assumptions of the Mokken model of monotone homogeneity (MMH; e.g., Sijtsma & Molenaar, 2002) by inspecting the H and H_g coefficients and by inspecting the item step response functions (ISRFs). Several methods have been proposed to check whether the ISRFs are monotonically increasing. In this study, we used the coefficient H_g for items ($g = 1, \ldots, k$) and coefficient H for a set of items. Increasing values of H and H_g between .30 and 1.00 (the maximum) mean that the evidence for monotone increasing ISRFs is more convincing, whereas values below .30 indicate violations of increasing ISRFs (for a discussion of these measures see Meijer & Baneke, 2003, or Sijtsma & Molenaar, 2002). Furthermore, weak scalability is obtained if $.30 \leq H < .40$, medium scalability if $.40 \leq H < .50$ and strong scalability if $.50 \leq H < 1$ (Sijtsma & Molenaar, 2002, pp. 60 - 61). However with personality data, it is very difficult to have items with $H_g \geq .4$. For that reason, we decided to set the lower bound to $H_g \geq .25$. Another reason for doing this was that otherwise there would be too few items per scale.

In addition to examining unidimensionality, we also checked to see if the facets of the Big Five were reasonably spread over the three scales, so that the three scales really measured something different from the Big Five These analyses resulted in an item bank consisting of 81 unique items spread over the three scales; Scale 1 “Cognitive Stress” consisted of 31 items, Scale 2 “Social Stress” of 27 items, and Scale 3 “Achievement Orientation” of 23 items. Tables 1, 2 and 3 show the H_g values and the item parameters, respectively for the three scales.

Item parameters. We estimated the item parameters using the graded response model (Samejima, 1969, 1997) and the computer program MULTILOG (Thissen, 2003) with marginal maximum likelihood estimation.

CAT Algorithm and Test Properties

Initial item selection. In this study we used ”the best guess” method (Parshal, et al., 2002). Because we wanted to prevent everyone from receiving the same first item, we choose the best five items of medium difficulty and programmed the application so that it randomly selected one of these five items.

Continued item selection. It is also important to decide how to select the next item in the CAT. In this study we used maximum Fisher information (MFI) as the criterion for selecting the next item. This item selection rule is commonly applied in CAT. The item with the highest amount of information at the examinee’s current trait level (θ) is selected (Parshal, et al., 2002).

θ estimation. After the examinee has responded to an item, the probability for that response is determined. The next step is to estimate θ. In this study, we used maximum likelihood estimation (MLE) to estimate θ, because there was no prior information available (Embretson & Reise, 2000; Parshal, et al., 2002).

Table 1. H_g Values and Item Parameters for the Cognitive Stress Scale

Item	H_g	a	b_1	b_2	b_3	b_4
1	.31	1.15	-2.84	-0.73	0.46	2.02
2	.30	1.08	-2.01	0.21	1.40	3.17
3	.37	1.75	-2.04	-0.87	-0.30	1.09
4	.38	2.05	-3.49	-2.28	-1.31	0.60
5	.27	0.91	-2.67	-0.07	1.09	3.40
6	.28	1.10	-4.32	-2.37	-1.19	0.62
7	.30	1.25	-2.66	-1.04	-0.20	1.73
8	.32	1.42	-3.22	-1.42	-0.56	1.07
9	.38	1.88	-2.02	-0.98	-0.35	0.81
10	.32	1.53	-3.24	-1.84	-0.95	0.47
11	.30	1.23	-4.06	-2.16	-0.87	1.38
12	.25	1.21	-3.79	-2.78	-1.84	-0.14
13	.25	0.83	-2.21	0.12	1.51	4.01
14	.34	1.44	-2.58	-0.72	0.08	1.60
15	.31	1.46	-5.13	-2.84	-1.91	0.46
16	.33	1.50	-4.10	-2.00	-0.71	1.19
17	.35	1.67	-3.22	-1.58	-0.79	1.14
18	.34	1.69	-2.87	-1.73	-1.09	0.50
19	.26	0.99	-3.80	-2.38	-0.66	1.66
20	.30	1.21	-3.87	-2.06	-0.55	1.53
21	.32	1.38	-2.66	-1.50	-0.76	0.72
22	.29	1.23	-4.76	-3.27	-1.70	0.60
23	.35	1.72	-3.05	-1.80	-1.12	0.43
24	.28	1.16	-3.14	-1.83	-0.35	1.38
25	.27	1.03	-2.56	-1.04	0.35	1.97
26	.27	1.04	-3.31	-1.28	-0.20	1.87
27	.29	1.05	-2.76	-1.28	-0.45	1.38
28	.33	1.39	-3.27	-2.11	-1.30	0.59
29	.27	1.03	-4.19	-1.64	-0.30	2.04
30	.29	1.24	-4.19	-2.44	-1.48	0.81
31	.34	1.78	-3.77	-2.54	-1.82	-0.07
Total	.31					

Table 2. H_g Value,
and Item Parameters for the Social Stress Scale

Item	H_g	a	b_1	b_2	b_3	b_4
1	.37	1.57	-3.21	-2.13	-1.24	0.51
2	.40	1.87	-1.68	-0.60	0.00	1.09
3	.29	0.97	-2.27	-0.56	0.78	3.22
4	.42	1.86	-2.46	-0.99	-0.27	1.48
5	.33	1.31	-3.63	-1.89	-0.87	1.06
6	.26	0.94	-5.37	-3.42	-1.04	1.66
7	.36	1.51	-3.43	-1.86	-0.98	0.76
8	.34	1.35	-2.99	-1.65	-0.52	0.92
9	.38	1.72	-2.04	-1.20	-0.56	0.53
10	.42	2.15	-1.60	-0.71	-0.09	1.00
11	.36	1.40	-1.67	-0.29	1.23	2.89
12	.37	1.64	-3.55	-2.38	-1.26	0.85
13	.40	1.95	-1.76	-0.88	-0.02	1.54
14	.33	1.32	-2.43	-0.82	-0.15	1.84
15	.29	0.95	-2.54	-0.32	0.78	3.18
16	.29	0.96	-3.10	-0.74	0.43	2.96
17	.31	1.25	-4.25	-2.01	-1.04	1.15
18	.39	1.75	-3.30	-1.71	-0.80	1.35
19	.31	1.19	-6.21	-3.95	-2.31	-0.07
20	.32	1.30	-4.73	-3.05	-1.11	1.46
21	.31	1.28	-5.30	-2.49	-1.00	1.30
22	.30	1.24	-7.24	-4.15	-2.34	0.50
23	.33	1.32	-3.17	-1.71	0.01	1.81
24	.27	0.90	-2.28	-0.30	0.68	3.05
25	.37	1.47	-2.92	-1.48	-0.69	0.99
26	.31	1.14	-5.84	-2.60	-1.45	1.12
27	.26	0.87	-2.28	-1.03	0.05	2.25
Total	.34					

Table 3. H_g Values and Item Parameters for Achievement Orientation Scale

Item	H_g	a	b_1	b_2	b_3	b_4
1	.29	1.25	-6.83	-3.27	-2.04	0.13
2	.36	1.69	-3.05	-1.18	-0.55	0.76
3	.30	1.11	-2.84	-0.92	0.03	2.12
4	.36	1.51	-2.46	-0.88	-0.01	1.77
5	.27	1.11	-3.26	-1.59	-0.72	1.31
6	.33	1.52	-3.43	-1.95	-1.27	0.47
7	.27	1.31	-3.99	-2.69	-1.67	0.00
8	.34	1.63	-2.56	-1.15	-0.59	0.41
9	.37	1.88	-2.80	-1.70	-0.81	0.59
10	.39	2.00	-2.13	-1.08	-0.39	0.78
11	.39	2.10	-2.66	-1.69	-0.86	0.74
12	.32	1.20	-2.95	-1.14	-0.01	1.68
13	.31	1.28	-3.47	-1.53	-0.57	1.19
14	.32	1.13	-2.59	-0.40	0.60	2.39
15	.36	1.42	-2.60	-0.82	-0.07	1.71
16	.35	1.45	-2.51	-0.97	-0.16	1.61
17	.28	1.18	-3.89	-1.99	-0.92	1.39
18	.28	1.17	-6.18	-2.96	-2.10	-0.17
19	.37	1.91	-3.45	-1.93	-1.12	0.31
20	.26	1.02	-4.92	-2.60	-1.25	1.87
21	.37	1.99	-2.70	-1.65	-0.87	0.74
22	.34	1.50	-4.28	-2.42	-1.53	0.83
23	.34	1.55	-3.16	-1.45	-0.73	0.91
Total	.33					

With MLE, the likelihood of a given response pattern is computed. The point along the latent trait continuum where this likelihood has a maximum is the value of θ. After each item response a new θ is estimated for the whole response pattern. After the new θ has been estimated, the search process for the next item is implemented. A problem with MLE is that no ML estimate can be obtained from all positive (category 4) and all negative (category 0) responses (Embretson & Reise, 2000). For that reason we decided to restrict the trait continuum from –4.4 through 4.8, so the item with the most information for that θ was then selected. This process was repeated until the stopping rule was reached.

Stopping rules. Simms and Clark (2005) specify four commonly used stopping rules:

1. A prespecified number of items has been administered,
2. The standard error (SE) of the trait estimate is below a prespecified value
3. The next item is less informative than a prespecified value,
4. A combination of these rules.

In this study we used a combination of these stopping rules. The most important stopping rule was that the test had to stop when SE < .32. We chose this value, because given that the classical reliability $\rho = 1 - \text{SE}^2$ for a standard normally distributed population, this value corresponds with $\rho \geq .9$ for each individual; this is excellent precision within classical test theory (Fliege, et al., 2005).

However analyses of the total test information functions (see Figure 1) showed that the test would never be able to reach SE < .32 for some people at very high levels of θ [1]. For that reason, we decided to use also another stopping rule: the next item had to give more information than .25. By using this stopping rule, an examinee does not have to complete all the items and be bothered with items that are not informative enough for his/her trait estimation. We also decided that the minimum number of items for each persons should be set at 9 items.

Results

Unidimensionality

Tables 1, 2 and 3 also show the H_g values for each item in every scale, and the H value for the total scale. The H_g values for the items in the Cognitive Stress scale varied from .25 to .38; the H value for the total scale was .31. For the Social Stress scale , the H_g values varied from .26 to .42 and the H value was .34 for the total scale. The Achievement Orientation scale had an H value of .33; the H_g values for the items varied from .26 to .39. According to Sijtsma and Molenaar (2002, pp. 60-61), the scales had weak scalability. However, it is very difficult for personality scales to have items with $H_g \geq .4$.

Item Parameters

Also included in Tables 1, 2, and 3 are the estimated item parameters for each item in the three scales. The threshold parameters (b_1, b_2, b_3 and b_4) varied between –S.13 and 4.01 for

[1] Note that SE = 1 / √I, where I is the total test information. In this case a SE of .32 equals a total test information of 10. There appears to be a problem within Multilog (Thissen, 1991), in that the total test information function I = 10 (left vertical axis) does not match with SE = .32 (right vertical axis). For that reason it is important to know whether the left vertical axis is correct and the right vertical axis is not.

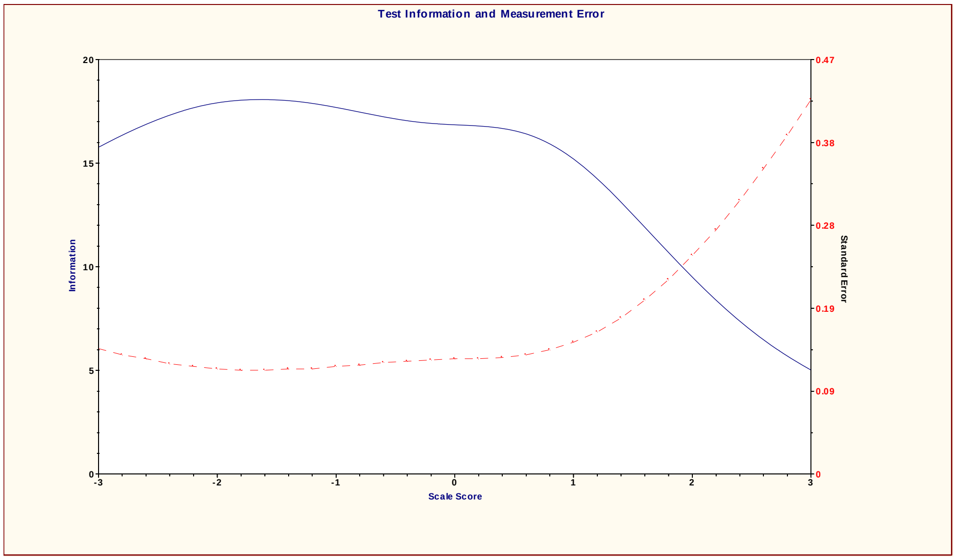
Cognitive Stress, between –7.24 and 3.22 for Social Stress and between –6.83 and 2.39 for Achievement Orientation. They covered a broad range, but the range was skewed to the left. The slope parameters (*a*) varied between .83 and 2.05 for Cognitive Stress, between .87 and 2.15 for Social Stress, and between 1.02 and 2.10 for Achievement Orientation.

Test Information Functions and Measurement Precision

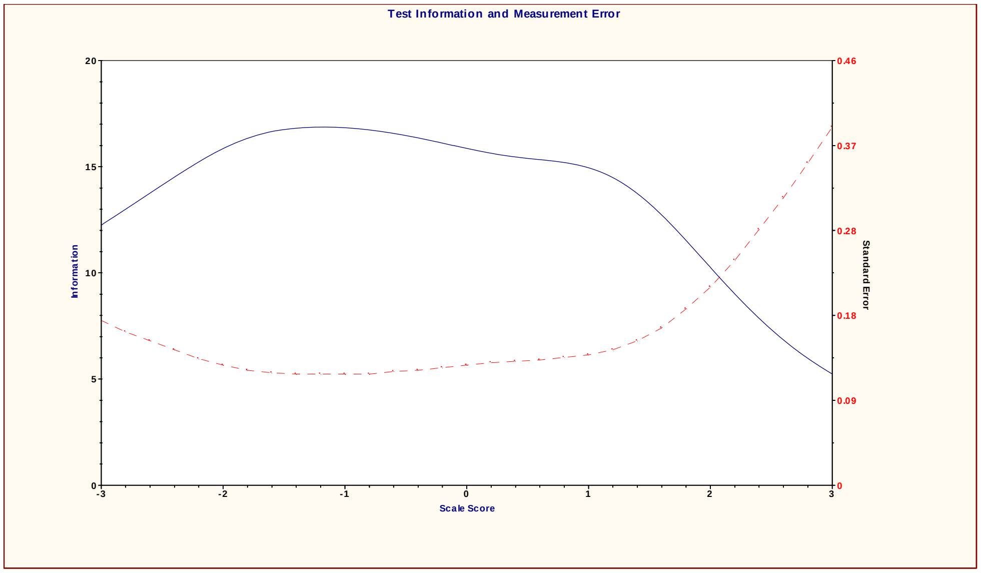
Figure 1 shows the test information functions for each scale. Each scale had high information for low θ values, but very low information for high θs. For the purpose of the test it is not a problem, because there is a reliable measurement for low θs, i.e. persons who are low on integrity and low reliability for persons of high integrity.

Figure 1. Total Test Information Functions

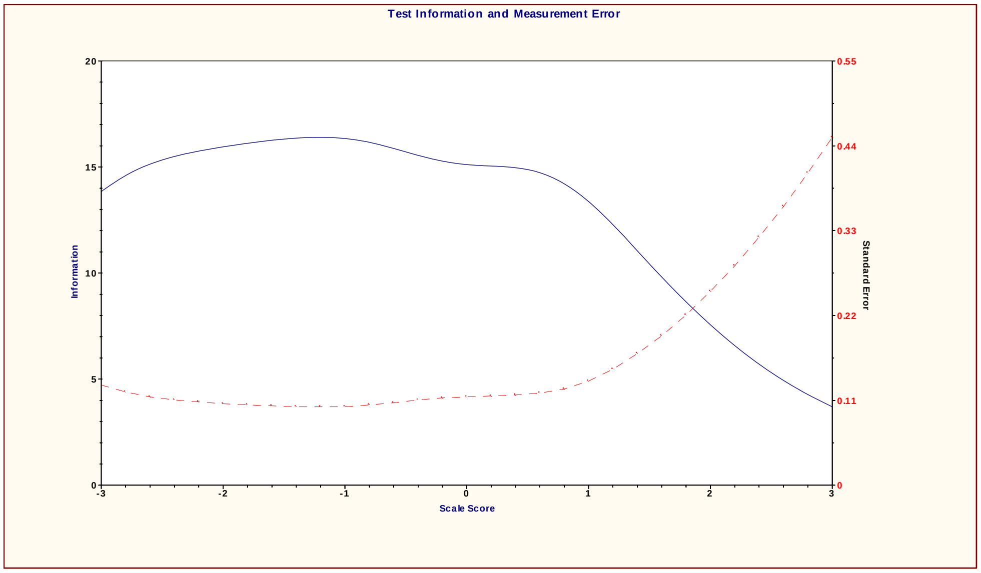
a. Cognitive Stress Scale



b. Social Stress Scale



c. Achievement Orientation Scale



STUDY 2

Method

Participants and Instruments

Data were collected from 271 participants with a mean age of 36.3 (SD = 12.0). There were 28% mostly White men and 72% mostly White women. 25.8% attended high school, 14.1% secondary vocational education, 36.9% had higher vocational education, and 23.2% had a university education. The participants completed the developed CAT for integrity, as described above. They also completed the WB5 (Schakel, Smid, & Jaganjac, 2007).

Procedure

A professional care organization cooperated in the research by recruiting participants using their Intranet. Soon it went by word of mouth. College students also participated in the research; participation was a compulsory part of the lecture on “Recruitment and Selection”.

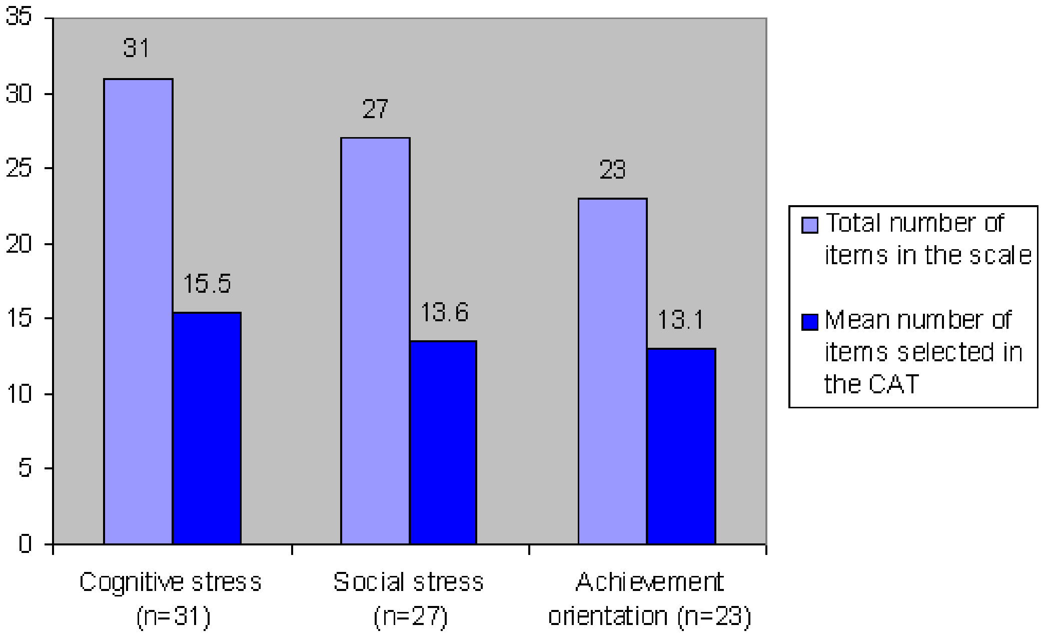
After participants applied to participate, they received an email with the link to the integrity test. As a reward they received, at least eight days later, a link to the WB5. After finishing this questionnaire they received their personal report in their mailbox. When the respondents filled out theWB5, they also answered the items that were used to develop the item banks for the integrity scales. By doing this it was possible to compare the score on the CAT scales with a full scale score based on the WB5 answers.

Results

CAT Efficiency

Figure 2 shows the mean number of items selected in the CAT in comparison with the total number of items in the scale. This figure shows a reduction of almost 50% for each scale.

Figure 2. Mean Number of Items Selected in the CAT Procedure



Items Selected and the *a* Parameter

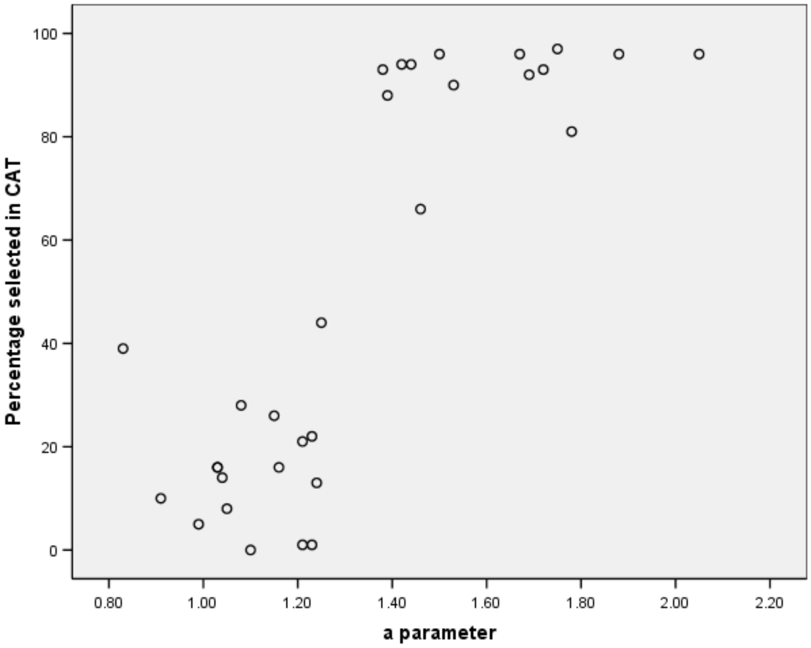
Table 4 shows for each item the *a* parameter and how many times (percentage) the item was selected in the CAT procedure. Figure 3 shows the relationship between these two variables. The figures show that the higher the *a* parameter, the more often an item was selected in the CAT procedure. For example, Item 4 from the cognitive stress scale “beliefs that he/she can handle problems” had the highest *a* parameter (a = 2.05) and 96% of the times this item was selected in the CAT procedure.

Table 4. *a* Parameters and Percentage of Times Selected in CAT for Each Item

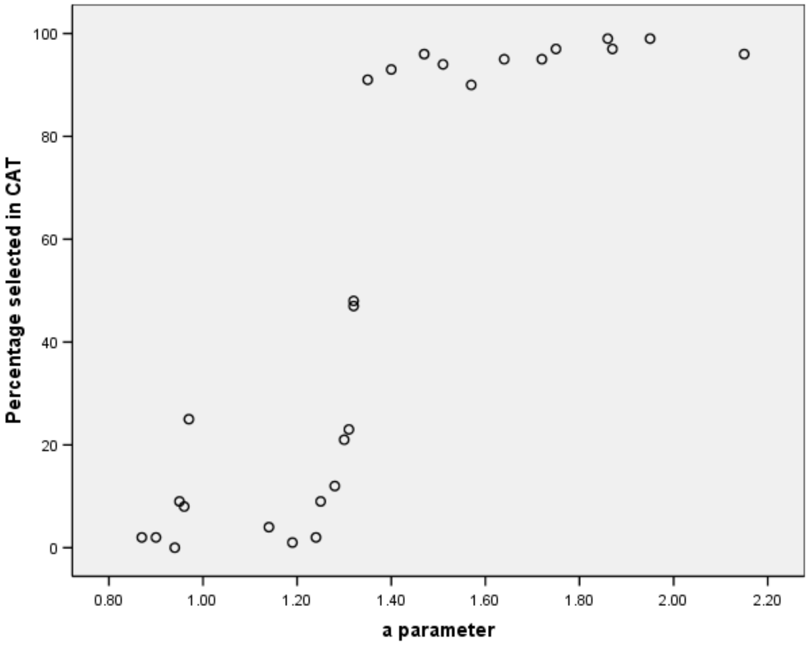
	Cognitive Stress		Social Stress		Achievement Orientation	
Item	Percent	*a*	Percent	*a*	Percent	*a*
1	26	1.15	90	1.57	0	1.25
2	28	1.08	97	1.87	94	1.69
3	97	1.75	25	0.97	17	1.11
4	96	2.05	99	1.86	100	1.51
5	10	0.91	23	1.31	9	1.11
6	0	1.10	0	0.94	90	1.52
7	44	1.25	94	1.51	0	1.31
8	94	1.42	91	1.35	92	1.63
9	96	1.88	95	1.72	94	1.88
10	90	1.53	96	2.15	96	2.00
11	22	1.23	93	1.40	98	2.10
12	1	1.21	95	1.64	19	1.20
13	39	0.83	99	1.95	18	1.28
14	94	1.44	47	1.32	37	1.13
15	66	1.46	9	0.95	81	1.42
16	96	1.50	8	0.96	97	1.45
17	96	1.67	9	1.25	16	1.18
18	92	1.69	97	1.75	0	1.17
19	5	0.99	1	1.19	94	1.91
20	21	1.21	21	1.30	13	1.02
21	93	1.38	12	1.28	96	1.99
22	1	1.23	2	1.24	53	1.50
23	93	1.72	48	1.32	95	1.55
24	16	1.16	2	0.90		
25	16	1.03	96	1.47		
26	14	1.04	4	1.14		
27	8	1.05	2	0.87		
28	88	1.39				
29	16	1.03				
30	13	1.24				
31	81	1.78				

Figure 3. Scatterplot of the a Parameters Against Percentage Selected in CAT

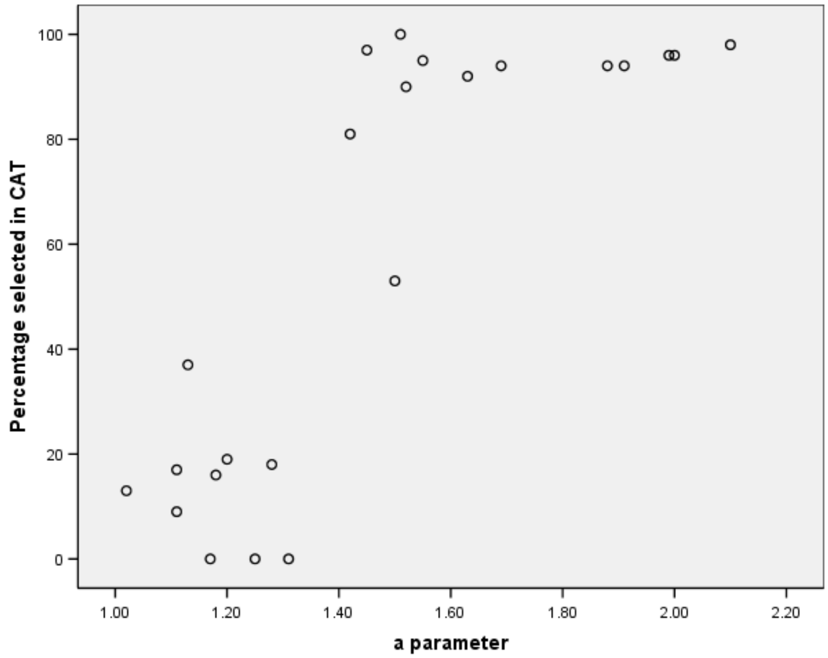
a. Cognitive Stress Scale



b. Social Stress Scale



c. Achievement Orientation Scale



CAT Versus the Full Scale

Table 5 shows the correlation between scores on the CAT and the full scale. There was a high correlation between these scores.

Table 5. Correlation Between CAT Scores and Full Scale Scores

Scale	Cognitive Stress	Social Stress	Achievement Orientation
Cognitive Stress (WB5)	**.845***	.589*	.311*
Social Stress (WB5)	.546*	**.899***	.161*
Achievement Orientation (WB5)	.403*	.172*	**.832***

*Significant at the .01 level (2-tailed).

Discussion

The objective of this study was to develop a CAT version of a personality-based integrity test that can be administered online, and can be applied for personnel selection. In addition, it was assessed whether adaptive administration of the test was an efficient way to do so.

The results showed that the CAT was more efficient than the conventional scale. With a CAT we could accomplish a reduction of items around 50%. With a mean of 42 items instead of 81 items, it was still reliable measurement. People also had nearly the same score if they answered all the 81 items or the CAT.

However the CAT procedure also had a drawback. Some items were almost never selected and some items were selected almost all the time. Figure 3 illustrates that this is due to the *a* parameter. Reise and Henson (2000) found similar results, and they also found that for most scales selecting the items with the highest *a* parameters produced similar results. This is why they suggested that it could be useful to select only the items that give most information and have the highest measurement precision.

More research is needed to investigate their suggestion. However, there are other issues that might need consideration. How about the content of the administered scales? Some items might dominate the scale, because they are selected for almost all the respondents. As a result, the scale might no longer, or to a lesser degree, measure what it is supposed to measure. Applying exposure controls (Sympson & Hetter, 1985, Van der Linden, & Veldkamp, 2004, 2007) could be an option. But exposure control also comes with a price. The CAT is forced to select items that are less informative. This implies that more items have to be selected to reach the same reliability level, and the CAT will be less efficient. Finally, another suggestion might be to write more "less informative" items with the same content. But that leads to one of the basic questions underlying personality scales: How many times can you ask people the same question over and over again? There are many issues related to the development of efficient personality-based CATs and more research is needed in this area. In this paper, we described the development of a CAT for measuring integrity. The instrument is now operational and this empirical date might enable us to further develop the instrument.

References

Cullen, M. J., & Sackett, P.R. (2003). Integrity testing in the workplace. In J. C. Thomas and M. Hersen (eds.), *Comprehensive handbook of psychological assessment: Volume 4. industrial and organizational assessment.* John Wiley and Sons Inc.

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

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B, Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, *14*, 2277-2291.

Meijer, R. R. & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354-368.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for windows. User's manual. Groningen*: The Netherlands: ProGamma.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the internet: New problems, old issues. *American Psychologist*, 59, 150-162.

Ones, D.S. (1993). The construct validity of integrity tests. Unpublished doctoral dissertation, University of Iowa City, Iowa.

Ones, D. S., & Viswesvaran, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale job applicant data sets. *Journal of Applied Psychology, 83*, 35-42.

Ones, D. S., & Viswesvaran, C. (2001). Integrity tests and other criterion-focused occupational personality scales (COPS) used in personnel selection. *International Journal of Selection and Assessment*, *9*, 31-39.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78* (4), 679-703.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Reid, J.E., (1967). *The Reid Report.* Chicago: John E. Reid and Associates.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment*, *7*, 347-364.

Sackett, P. R., Burris, L. R., & Callahan, C. (1989). Integrity testing for personnel selection: Un update. *Personnel Psychology, 42*, 491-529.

Samejima, F. (1969). Estimation of latent 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 items response theory* (pp. 85-100). New York: Springer-Verlag.

Schakel, L., Smid, N. G., & Jaganjac, A. (2007). *Workplace Big Five professional manual.* Utrecht, The Netherlands: PiCompany B.V.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA. Sage.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the schedule for nonadaptive and adaptive personality (SNAP). *Psychological Assessment, 17,* 28-43.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. Proceedings of the 27th Annual Meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thissen, D. (2003). Multilog for Windows (Version 7.0) [Computer software]. Lincolnwood, IL: Scientific Software International.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273-291.

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398-417.

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.

Wanek, J. E. (1999). Integrity and honesty testing: What do we know? How do we use it? *International Journal of Selection and Assessment, 7*, 183-195.