Personnel Assessment and Decisions

Volume 2 | Issue 1 Article 3

2016

Cheating on Unproctored Internet Intelligence Tests: Strategies and Effects

Wim Bloemers
Open Universiteit Nederland, wim.bloemers@ou.nl

Arjan Oud
Oud Career planning, info@arjanoud.nl

Karen van Dam
Open Universiteit Nederland, Karen.vanDam@ou.nl

Recommended Citation

Bloemers, Wim; Oud, Arjan; and Dam, Karen van (2016) "Cheating on Unproctored Internet Intelligence Tests: Strategies and Effects," *Personnel Assessment and Decisions*: Vol. 2 : Iss. 1 , Article 3.
DOI: https://doi.org/10.25035/pad.2016.003
Available at: https://scholarworks.bgsu.edu/pad/vol2/iss1/3

Cheating on Unproctored Internet Intelligence Tests: Strategies and Effects

Wim Bloemers[1], Arjan Oud[2], and Karen van Dam[1]

1. Open Universiteit Nederland
2. Oud Career Planning

ABSTRACT

A crucial issue concerning unproctored Internet-based testing (UIT) of cognitive ability is its susceptibility to cheating. Whereas evidence indicates that cheating during UIT occurs, there is still little information about possible cheating strategies and their effects on (sub)test performance. Using a randomized experimental design, this study investigated the direct effects of cheating on an Internet-based test of cognitive ability by comparing test performance of cheaters (participants who were instructed to cheat) and successful cheaters (participants who thought their cheating had been successful) with that of noncheaters. Successful cheaters obtained substantially higher scores compared to cheaters who thought they had been unsuccessful in cheating and noncheaters. The effect of cheating depended on subtest type and the number and type of cheating strategy being used. Suggestions are made for further research and for safeguarding future UIT procedures from cheating.

KEYWORDS

unproctored Internet testing, cognitive ability test, cheating, experiment

Today, anyone can complete cognitive ability tests anywhere in the world, at any time, granted that his or her computer is linked to the Internet. This unproctored Internet-based testing (UIT) of cognitive ability, or *controlled delivery*, does not have human supervision during the test procedure. A login and ID are provided to the test taker but no further identification is required (for a recent overview of UIT aspects, see Scott & Lezotte, 2012).

Compared to more conventional ways of testing, UIT has several advantages, such as more efficiency, high-tech image, and greater flexibility (Arthur, Glaze, Villado, & Taylor, 2010; Barak, 2010; Lievens & Burke, 2011; Tippins, 2009). Business case examples (e.g., Gibby, Ispas, McCloy, & Biga, 2009; Kaminsky & Hemingway, 2009) explain how organizations can benefit from UIT implementation. Survey results from Ryan et al. (2015) show that, among companies who use computerized testing, 40% uses UIT for all their test procedures, including high stakes, whereas only 20% still uses a supervised format. Main reasons for this increase in UIT is demonstrated value and efficiency, fairness, more convenience for applicants and hiring managers, cost effectiveness, larger applicant pools, reduction of hiring time, and increased precision in measurement through the use of CAT and IRT (Ryan et al., 2015; Scott & Lezotte, 2012).

Notwithstanding all these positive aspects of UIT, concerns have been raised relating to disadvantages of UIT, including problems with the Internet connection and the impersonal nature of remote computerized testing (Tippins et al., 2006). By far the most severe objection against UIT concerns its vulnerability to cheating (e.g., Arthur et al., 2010; Tippins et al., 2006). UIT is not monitored by a human test administrator. In a high stakes situation, test takers have the opportunity and a motivation to raise their score by using cheating strategies (Duffield & Grabosky, 2001; Tippins et al., 2006). Given the lack of control and the importance of applicants' test scores for hiring decisions, cheating on UIT appears a rather obvious way to pass this selection hurdle in times when jobs are highly valued. However, organizations in general are rather optimistic about the psychometric integrity of UIT procedures. About 50% estimate the percentage of frauds on a UIT below 10%, and only about 10% of the organizations expect that *more* than 10% of the applicants will cheat (Ryan et al., 2015). Research based estimates of people cheating on web based tests vary from 7 –50% (Arthur et al. 2010). In a Cubiks study (2006) about 10% of the test takers admitted to have cheated.

In line with the increased use of UIT and concerns about cheating, research on cheating has increased as well.

Corresponding author:
Wim Bloemers
Overtoom 163 D 1054 HG, Amsterdam, The Netherlands
Email: Wim.Bloemers@ou.nl
Phone: 00-31-6-11175592

Yet, little is known about the effects of cheating in general and how much the results on different subtests are affected by cheating. Moreover, we know very little about strategies people use when trying to cheat and their possible effects.

The objective of our study was to investigate the impact of cheating efforts on the outcomes of a cognitive ability test battery and determine possible differential effects for the various subtests and cheating strategies. We used a randomized two-group experimental design to compare the scores of cheaters (research participants who were instructed to cheat) with those of noncheaters. This design allows for a more thorough assessment of cheating effectiveness than the indirect score change evaluation procedure that has been used in previous studies (e.g., Arthur et al., 2009).

Theoretical Background

UIT or controlled delivery implies several threats to test reliability and validity. First, unproctored test events generally lack standardization; variations in responses might occur owing to factors outside the candidate, such as noise and distractions, performance of the test taker's computer, and quality of the Internet connection (Potosky & Bobko, 2004).

Another potential threat to the validity of UIT is cheating. Cheating on maximum measures can be defined as intentionally using any means, whatsoever, to produce an answer on an item that does not represent the true position of the candidate on the underlying latent variable. Examples are the use of unauthorized reference materials and consulting others (Landers & Sackett, 2012).

Especially under "high stakes" conditions, where scores are used for hiring or other employment decisions, candidates might be tempted to cheat with strategies like using a calculator, a dictionary, the Internet, test preparation, the help of others, or even having another person take the test. The absence of proctoring and the high stakes situation represent the factors opportunity and pressure of the *fraud triangle* (Cressey, 1973). The fraud triangle consists of three elements: pressure to cheat, opportunity to cheat, and rationalization. It was developed by Cressey to explain *why* seemingly honest people commit fraud, and it thus gives an explanation of *why* people might be tempted to cheat. Our experiment includes both the factors opportunity and pressure to cheat.

Although cheating on a UIT might seem easy at first sight, sophisticated test features can limit cheating effectiveness. Owing to fixed item time limitations in speeded tests, increasing item difficulty, and the use of adaptive test technology, effective cheating on unproctored Internet-based tests of cognitive ability might be quite difficult. Findings indeed indicate that unproctored test situations do not necessarily lead to increased test scores (Lievens & Burke, 2011; Shepherd, Do, & Drasgow, 2003; Tippins et al., 2006), which might also imply that cheating occurred that was not successful.

As Bartram (2005) notes, the essential question is not so much *whether* candidates cheat but rather *how* cheating efforts affect test outcomes. This points to three aspects: how do people cheat, what is the effect of cheating on different subtests and total test score, and should cheating be considered a serious threat for the further development and implementation of UIT? (Landers & Sackett, 2012; Tippins, 2009; Tippins et al., 2006).

The Present Study

To investigate the effect of cheating on a UIT of cognitive ability, we conducted an experiment using a randomized, two-group design and a web-based speeded test for cognitive ability. Cheating was defined as any conscious attempt to achieve the highest possible test score through the use of inappropriate or fraudulent means, such as the use of aids (e.g., calculator, dictionary, Internet), manipulating the procedure, receiving help from others, or foreknowledge (i.e. having had access to test content prior to the assessment; Scott & Lezotte, 2012). Based on previous studies (Hausknecht, Halpert, Di Paolo, & Moriarty-Gerrard, 2007; Nye, Do, Drasgow, & Fine, 2008), it was expected that cheating would result in higher test scores (*Hypothesis 1*) and that this effect would depend on the number of cheating strategies used (*Hypothesis 2*). As different subtests require different cheating means and our test consisted of seven different subtests, we expected the largest cheating effect for candidates using a number of cheating strategies. Using help from others and technical manipulation (sabotaging the web-based test application to omit time constraints, manipulating the test interface, or any other technical intervention to obtain a higher score) can be seen as "meta" strategies, because these two strategies enable use of all other strategies as well. The effectiveness of specific strategies was investigated.

METHOD

Design and Procedure

Using a randomized two-group experimental design, participants were randomly assigned to either a control group, receiving no specific instructions, or a "cheating" group, who completed the test after explicit instructions to cheat and maximize their test score. The cheating instruction emphasized that participants could do anything in order to increase their score. To further stimulate participants to cheat, two cash prizes of €100 (about $110) were made available for the two participants in the cheating group with the highest scores, thus creating a proxy for a high-stakes condition. In the control group, two prizes of €100 would be raffled.

Participants in the cheating condition received the cheating instruction a week in advance of the log-in codes to ensure that they would have a similar opportunity to prepare themselves as in a real selection procedure. As participants knew from the instruction which test would be used, they were sufficiently able to prepare themselves. To remind them of the importance of cheating, the cheating instruction was repeated upon logging in. After completion of the test, all participants had to indicate whether they had tried to cheat and, if so, which strategies they had used. Finally, participants were debriefed and received feedback on their scores.

Participants

Participants were (former) students of the Open University in the Netherlands. The Open University provides distance learning, enabling students to study at times and in places that suit them. Through e-mail, students were invited to participate in the study. Upon a positive response, participants received a log-in code for the test and a deadline for completing the test.

Of the 5,231 students who had been approached, 1,015 (22%) agreed to participate in the study. A total of 463 participants (46%) completed all tests, 255 in the control group and 208 in the cheating group. Seven participants in the control group still indicated to have cheated and were excluded from the study, leaving 248 participants in this group. Similarly, 30 participants were removed from the cheating group because they indicated to have refrained from cheating, leaving 178 participants in this group. The groups did not differ in age (M = 38.3, SD = 8.78), gender (80% female), and educational background; participants had completed either lower vocational education (25%), higher vocational education (45%), or university (29%).

Measures

Cognitive ability. The cognitive ability test used in this study was the Q1000 online test of Meurs HRM, a major supplier of web-based tests and instruments in the Netherlands (for a brief test description and item examples in Dutch, see: https://cdn.q1000.nl/gebruikersinformatie/cognitieve-capaciteiten-gebruikershandleiding.pdf. The Q1000 is widely used in the Netherlands and has shown adequate reliability and validity in most studies (Evers, Lucassen, Meijer, & Sijtsma, 2009). The test used in this experiment was the shortened version of the Q1000, containing 68 items divided over seven subscales; two subscales addressed numerical abilities, Calculations (8 items) and Number series (8 items); two subscales addressed visual perception, Figures (10 items) and Cubes (6 items); and three scales addressed verbal abilities, Analogies (13 items), Syllogisms (8 items), and Vocabulary (15 items). With seven subscales we had sufficient variation in our item pool to elicit different cheating strategies. Also, items differentiated between more *g*-loaded fluid tasks, such as verbal reasoning and figure completion series, and more crystallized tasks, like calculations and vocabulary. The idea was that the total test would elicit and facilitate various cheating strategies. The test was timed, with a fixed time for each subtest item.

Cheating. Participants were asked whether they had tried to cheat when taking the test (0 = *no*; 1 = *yes*); if so, whether they thought they had been successful in cheating (0 = *no*; 1 = *yes*); and how difficult it had been to cheat (1 = *very easy*; 5 = *very difficult*). Participants were also invited to comment on their cheating behavior (open answer format).

Cheating strategies. Participants could indicate which strategies they had used by ticking one or more of the cheating strategies that were listed (0 = *no*; 1 = *yes*), calculator, dictionary, Internet, help of others, test books, technical manipulation, foreknowledge (having specific knowledge of the test content), and other strategies (open-answer format).

Cheating expectation. Participants were finally asked whether they expected they would cheat on an unproctored test in a real-life selection situation.

Background variables. Participants provided data on gender, age, and educational level.

RESULTS

Before testing the hypotheses, we investigated the internal consistency estimates (Cronbach's alpha) of the subscales for the two groups separately and the group as a whole. As the findings in Table 1 show, subscale reliabilities of the cheating and control condition were comparable. The reliability of the aggregated cognitive ability measure was .83 in the cheating condition and .78 in the control condition. Together, these findings indicate that participants' cheating efforts did not affect test reliability. In addition to Cronbach's alpha, we computed Spearman-Brown split-halves reliabilities by correlating odd–even items, as alphas are less appropriate for speeded tests (Allen & Yen, 2002). The split halves were of the same magnitude as the alphas, respectively .82, .80, and .81 for the aggregated test scores in the fraud-, control- and total group.

Cheating and Test Performance

With an ANCOVA the average total test scores of the "honest" and the "cheating" group were compared to determine the impact of cheating. Because previous research shows that there is a solid correlation between educational level and intelligence test performance (about .50, Neisser et al., 1996), and preliminary analyses indeed indicated that education was positively related to test performance, this variable served as a covariate in the analysis.

TABLE 1.
Internal Consistency Reliability Estimates[1] of the Subtests

	Number of items in test	Control group (n = 253)	Cheating group (n = 178)	Both groups (N = 431)
		α	α	α
Calculations	8	.41	.40	.42
Numbers	8	.56	.59	.58
Figures	10	.41	.46	.44
Cubes	6	.51	.48	.50
Analogies	13	.53	.47	.50
Syllogisms	8	.50	.52	.51
Vocabulary	15	.61	.65	.64
Full test	68	.78	.83	.81

Note. [1] Cronbach's alpha.

There was a significant difference between the "honest group" (M = 34.64; SD = 8.01) and the "cheating group" (M = 38.14; SD = 8.59); F(1, 416) = 17.94, p < .001, d = .40); the scores of the "cheating" group were higher. The impact score of .40 indicates a medium effect size (Cohen, 1988). The effect of the covariant "education" was significant (F(1, 416) = 46.516, p < .001), indicating that educational level had an effect on the test scores.

The majority of the participants in the cheating condition indicated that cheating had not been an easy task. Only 18% considered cheating easy, whereas 71% indicated that cheating had been difficult. Moreover, only 58% of the cheaters thought they had been successful in cheating. Based on these self-reports, we decided to split the group of cheaters into (self-rated) ineffective cheaters (n = 75) and effective cheaters (n = 103), and compare their performance with test performance of the control group (n = 248). To establish the impact of (no, ineffective, and effective) cheating on general test performance, an ANCOVA was conducted, with education again as a covariate. Table 2 presents the outcomes of this analysis.

In general, cheaters' assessment of their cheating effectiveness appeared correct; on average, participants who reported that their cheating had been successful had significantly higher total test scores (M = 39.70, SD = 8.41) than either participants who considered their cheating unsuccessful (M = 35.95, SD = 8.43) or participants in the control group (M = 34.64, SD = 8.01). Whereas the latter two groups did not differ in test performance, the effect score (Cohen's d) between successful cheaters and the other two groups was .59. Total mean test time was 1538.11s (SD = 415.90) for the successful cheat group, 1516.30s (SD = 396.73) for the unsuccessful cheat group, and 1625.94 s (SD = 346.99) for the control group. The time difference between the control group and both the unsuccessful and successful cheaters was significant (p < .022 and p < .006).

TABLE 2.
Impact of Cheating Efforts on Test Performance

	Control group	Cheating group			
	(n = 248)	Ineffective (n = 75)	Effective (n = 103)		
	M (SD)	*M (SD)*	*M (SD)*	*F*	Cohen's d^2
Calculations	3.80^a (1.65)	4.01^a (1.57)	4.54 (1.70)	7.45 **	.42
Numbers	3.34^a (1.76)	3.41^a (1.82)	4.19 (1.70)	8.85 ***	.47
Figures	4.04^a (1.60)	4.41^a (1.47)	4.40^a (1.68)	2.61	.17
Cubes	3.21^a (1.62)	3.41^a (1.49)	3.86 (1.53)	6.20 **	.38
Analogies	7.60^a (2.28)	7.58^a (2.18)	8.08 (2.13)	1.86	.22
Syllogisms	4.35^a (1.64)	4.37^a (1.51)	4.67^a (1.69)	1.43	.19
Vocabulary	8.29^a (2.91)	8.75^a (2.97)	9.95 (2.79)	12.09 ***	.54
Full test	34.64^a (8.01)	35.95^a (8.43)	39.70 (8.41)	13.95 ***	.59

Note. [a] indicates similarity of group means; means that do not share any subscript are significantly different at p < .05.
[b] Cohen's d relates to score differences of the effective cheaters in comparison to the other participants (ineffective cheaters and control group). ** p < .01. *** p < .001.

To establish how cheating had impacted the subtest scores, first, a MANCOVA was conducted using the scores of all subtests as dependent variables and the three groups as comparison factor. The outcomes indicated a significant difference in at least one subtest (*F* (7, 406) = 3.59, *p* < .001). Subsequent ANCOVAs for the subtests separately showed that the effective cheaters had done significantly better than the other two groups on four of the seven subtests: Calculations (*M* = 4.54 vs. *M* = 4.01 and *M* = 3.80; *p* < .01); Numbers (*M* = 4.19 vs. *M* = 3.41 and *M* = 3.34; p < .001); Cubes (*M* = 3.86 vs. *M* = 3.41 and *M* = 3.21; *p* < .01), and Vocabulary (*M* = 9.95 vs. *M* = 8.75 and *M* = 8.29; *p* < .001). For Vocabulary (*d* = .54), Number series (*d* = .47), and Calculations (*d* = .42), effect sizes (Cohen's *d* (1988)) were substantial. For Cubes the effect size *d* was .38. The groups did not differ significantly in their scores for the subtests Figures, Analogies, and Syllogisms.

Cheating Strategies

On average, cheaters used 1.84 (*SD* = .94) cheating strategies; and in this respect ineffective and effective cheaters did not differ (*M* = 1.82, *SD* = .87; vs. *M* = 1.85, *SD* = .98; *t* (177) = -.23, *ns*). The calculator appeared the most popular cheating strategy (*n* = 135), followed by help from others (*n* = 87), a dictionary (*n* = 55), and the Internet (*n* = 55). Cheaters used technical manipulation (e.g. trying to sabotage the test session for a second chance; *n* = 12), foreknowledge (*n* = 25), and test books (*n* = 1) less frequently.

The number of strategies used was positively related to overall test performance (*r* = .26, *p* < .001). To establish the effectiveness of the different strategies, independent *t*-tests were conducted comparing the overall test scores of participants who either had or had not used a certain cheating strategy. As the findings in Table 3 show, all strategies except foreknowledge resulted in significant effects, with *p* values varying according to the number of participants using a specific strategy.

We also checked for differential strategy use between successful and unsuccessful cheaters. Two strategies showed differential use: Successful cheaters used about twice as much *help from others* (75.9% vs. 38.0%) and *technical manipulation* (9.3% vs. 5.0%). There were no obvious differences for other strategies.

A multiple regression analysis with the specific strategies as predictors showed the importance of using a dictionary (β = .15, *p* < .01) and help from others (β = .14, *p* < .01) for overall test performance. Together, the use of cheating strategies explained 8% percent of the variance in overall test performance.

Concerning future cheating, participants generally reported low cheating expectations. Only 7% of the participants in the cheating group expected they would (possibly) cheat in a real-life selection procedure; 43% expected this chance to be small; and 50% indicated they had absolutely no intention to cheat. This expectation was unrelated to cheating effectiveness (*r* = -.05, *ns*) and cheating difficulty (*r* = .09, *ns*).

TABLE 3.
Impact of Cheating Strategy on Overall Test Performance

		No	Yes	*t* (421)	Cohen's *d* [a]
Calculator	*M*	34.82	38.71	4.54 ***	.47
	SD	8.50	7.59		
	n	288	135		
Dictionary	*M*	35.37	40.69	4.48 ***	.65
	SD	8.37	7.20		
	n	368	55		
Internet	*M*	35.85	39.31	2.01 *	.41
	SD	8.45	7.10		
	n	368	55		
Help from others	*M*	35.22	39.29	4.06 ***	.49
	SD	8.22	8.38		
	n	336	87		
Technical manipulation	*M*	35.93	40.98	1.96 *	.60
	SD	8.26	12.48		
	n	411	12		
Foreknowledge	*M*	36.09	34.70	.61	-.17
	SD	8.35	11.00		
	n	412	11		

Note. [a] the formula for pooled SD was used. ** *p* < .01. *** *p* < .001.

DISCUSSION

The increased popularity of UIT has raised concerns about its psychometric integrity. Absence of a human test administrator monitoring the test environment can easily lead to cheating and manipulated test scores (Ployhart, Weekley, Holtz, & Kemp, 2003). Although previous research (e.g., Nye et al., 2008) indicates that cheating in an unproctored test environment occurs, the effectiveness of different cheating strategies has not been established yet. Therefore, the aim of the present study was to investigate how people cheat and what effect cheating had on subtest scores and the total test score by explicitly inviting test takers to cheat on a UIT of cognitive ability.

The hypothesis that cheating would pay off was partly confirmed. Although the findings indicated that cheating on an UIT of cognitive ability results in enhanced test performance, they also showed that cheating efforts paid off for some but not for others. About 40% of those who had tried to cheat reported that they found cheating difficult and doubted whether they had been successful. Their perceptions appeared to be right; on average, they obtained a test score similar to the control group that had not been instructed to cheat, whereas the other 60% of the cheaters obtained a much higher score. This raises the question of why some cheaters were less effective than others. In the open answer space, cheaters complained about the timed character of the test that had put a limit to their cheating efforts. Others indicated that they had not prepared themselves well enough to be able to cheat effectively. More generally, these accounts indicate that test takers might be able to improve cheating effectiveness only by thorough preparation.

Yet, cheating paid off for the other cheaters. Test performance of the effective cheaters was on average higher than test performance of the ineffective cheaters and the control group. The size of this effect varied with subtest. Cheating appeared most effective for the Vocabulary, Numbers, and Calculations subtests, on which performance can be easily enhanced through the use of a dictionary or a calculator whether or not combined with help from others. Cheating did not affect test performance on the subscales Analogies, Figures, and Syllogisms. Apparently, the available cheating strategies were not adequate for enhancing test performance on these subscales, not even with technical manipulation and help of others. A possible explanation is that these items require language comprehension and complex reasoning, which takes a lot of time and requires substantial cognitive capacity. From the point of cheating prevention, it could be advised to construct and use unproctored test batteries consisting of items with a high g load (complex reasoning).

With an effect size of .59 for overall test performance, the effect of cheating was substantial for the group of effective cheaters, supporting other studies that similarly detected cheating effects (Arthur et al., 2010; Nye et al., 2008). Moreover, it should be noted that the six highest scores were all found among the group of effective cheaters, implying that if a top-down selection strategy would be used, cheaters instead of honest candidates would be hired. For example, in case of a selection rate of 10%, 28 of the 43 selected candidates (65%) would belong to the fraud group, whereas only 18 fraud candidates were expected based on group ratios. However, one should be careful with generalizing these numbers to a real life UIT, where the percentage of frauds will be far less than in our forced fraud experiment. Regarding the prevalence of frauds in real life UIT procedures, it is notable that seven subjects (3%) were removed from the initial honest group because they admitted still to have cheated. Another indication of "real" die-hard frauds is the percentage of subjects indicating they would cheat in a real-life test situation if they had the chance. About 7% of the subjects in the cheat group indicated they surely or probably would cheat in a real life UIT situation. Remember that in the Cubiks study (2006) about 10% frauds were found. Together, this evidence indicates that about 7–10% of candidates on a UIT in a high stakes situation might be expected to cheat, with a lower bound of 3%. Organizations that expect that no more than 10% of the candidates will cheat (Ryan et al., 2015) are probably right.

Most cheaters used a combination of cheating strategies. This appeared effective, because test performance was, albeit modestly, related to the number of cheating strategies that were used, thus supporting our second hypothesis. Type of strategy appeared important for cheating effectiveness; in particular, the use of a dictionary, calculator, and/or combined with help from others was related to enhanced test scores. It should be noted that the (large) effect of the use of a dictionary on the total test score is overestimated due to the relatively large number of items in the vocabulary subtest. If used creatively, technical manipulation also can have a very large effect. Relying on technical manipulation, one participant, for example, indicated that he had photographed the items of the test, then pulled the plug and requested a new login code, claiming his Internet connection had failed. This candidate obtained the highest score of all, thus "earning" the €100 prize. Another one routinely copied each screen, closed the test window, solved the copied item, logged in again and typed in the presumed correct answer. Technical manipulation allows candidates to solve items outside the fixed time frame. On further inspection, four of the six high scorers had impossible low total test time scores, well below 1000 s, pointing to technical manipulation. In a real selection situation, these candidates should be required to take a proctored verification test. Impossible low time scores are also an explanation why mean total time in the successful and unsuccessful cheat groups was lower. In the control group, subjects used more time because of the power character of the subtests requiring all their effort and time, whereas in the unsuccessful cheat group, a relatively large group of people may have given up or were forced to guess because of cheat failure. In the successful cheat group, the number of impossible low time scores, due to technical manipulation, reduced mean total test time.

Study Limitations

The findings of this study should be interpreted in the light of several limitations. First, this study used an experimental situation instead of a high stakes real-life selection situation. Although we tried to create a high-stakes condition by promising financial rewards for high performance, the outcomes in real selection situations might be more powerful, such that cheaters might show more thorough and creative efforts to increase their test scores than the subjects in our cheating condition (see Drasgow, Nye, Guo, & Tay, 2009). Second, our between-subjects design might have led to an underestimation of the effect sizes owing to increased error associated with comparing two different groups. Although a within-subjects design has the advantage of greater statistical power (Greenwald, 1976), it might also result in practice effects (Arthur et al., 2009; Lievens & Burke, 2011). Underestimation of the real cheating effect has also occurred if some participants in the control group have cheated. Although control-group participants who had admitted to have cheated were removed from the data set, others may have cheated as well without reporting their cheating behavior, thus staying undetected. Owing to these factors, the actual effect of cheating might be larger than our findings revealed.

Together our findings indicate that people who manage to cheat are likely to raise their test score substantially. In the literature, different measures for detecting and decreasing cheating have been mentioned, including proctored test-taker authentication, increasing perceptions of accountability, keystroke analysis, adaptive testing, and using speeded tests (e.g., Foster, 2009; Gibby et al., 2009; Lerner & Tetlock, 1999). Future research might focus on the effect of these different precautions on cheating efforts, strategy selection, and strategy effectiveness. The two "meta" strategies, help from others and technical manipulation, seem the most effective. They were used twice as much in the successful fraud group compared to the unsuccessful fraud group. But they are also very hard to apply effectively. Research is also needed to establish why some cheaters were more effective than others, relating cheating effectiveness to individual characteristics such as creativity, openness to experience, conscientiousness, or emotional stability and investigate the impact of cheating preparation. Creative people cheat more easily and with greater effect, so this seems a fruitful perspective on the person side of cheating (Gino & Ariely, 2011). It also points to *rationalization* as the third element of the Fraud Triangle (Cressey, 1973), because creative people can invent successful fraud strategies and sophisticated rationalizations as well.

Additionally, future research might examine which strategies are effective for which specific subtests, thus explaining why some subtests are more and others are less vulnerable to cheating. Based on the outcomes of our study, specific preventive measures can be taken. Although the reliability of the scores was not affected by cheating, on the individual test takers level cheating paid off: the six highest scores belonged to the cheating group. Four of them used a technical manipulation. Thus, although test scores can show sufficient reliability, cheating can profoundly distort the ranking of test outcomes, affecting both fairness and predictive validity.

A Scenario for Preventing Fraud on UIT

From our experiment it becomes undoubtedly clear that any UIT procedure that is not technically "sound," will be contaminated by cheating behavior. Preventing technical manipulation thus is the first and most important measure that should be taken regarding the psychometric integrity and fairness of UIT. Attempts to omit time limits, impossible low time scores, and requests for a new login should automatically lead to a removal from the current UIT-procedure, eventually followed by a proctored verification test. Keystroke dynamics and response latency can be very effective fraud deterrents (Scott & Lezotte, 2012). Limited test time proved effective in preventing cheating, but caution should be taken not to make a procedure too applicant unfriendly by severe fixed time limits thereby repelling favorable candidates beforehand.

A possible behavioral preventive measure is candidates making sign an honesty statement (Dwight & Donovan, 2003; Fan et al., 2012). Web cam monitoring can prevent candidates from using help of others but can deter promising candidates because of privacy invasion.

The above measures, combined with adaptive (CAT), IRT test technology and a large item bank consisting of items with a high *g*-load, minimize the chance that cheating attempts will be successful and that cheaters will be selected in, owing to their cheating efforts. When in doubt about the test performance of a candidate, one should *always* require a proctored verification test different from the first UIT version.

In the introduction we mentioned three crucial aspects of our research: How do people cheat, what is the effect of cheating on different subtests and total test score, and should cheating be considered a serious threat for the further development and implementation of UIT? We now know that cheating, as a general phenomenon, is probably not a serious threat for UIT; at the same time, successful individual cheating strategies, such as technical manipulation, can have a large impact and cause severe damage to the fairness of a selection procedure. When organizations and assessment psychologists thoroughly communicate with candidates, being transparent and explaining the above safety measures from a fairness and validity point of view, UIT can be a very powerful, efficient, fair, and candidate-friendly mechanism in a selection procedure, taking personnel selection to a higher professional level.

REFERENCES

Allen, M. J. & Yen, W.M. (2002). *Introduction to measurement theory.* Prospect Heights, IL: Waveland Press.

Arthur, W. Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2009). Unproctored internet-based tests of cognitive ability and personality: Magnitude of cheating and response distortion. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 39-45.

Arthur, W. Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored Internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18*, 1-16.

Barak, A. (2010). Internet-based psychological testing and assessment. In R. Kraus, G. Stricker, & C. Speyer (Eds.), *Online counseling: A handbook for mental health professionals* (2nd ed, pp. 225-256). London, UK: Academic Press.

Bartram, D. (2005). The changing face of testing. *The Psychologist*, 18, 666-668.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York, NY: Academic Press.

Cressey, D. R. (1973). *Other people's money: A study in the social psychology of embezzlement.* Montclair, NJ: Patterson Smith.

Cubiks (2006). *Review of candidate attitudes towards unsupervised computer-based testing.* Retrieved from http://www.cubiks.com.

Drasgow, F., Nye, C. D., Guo, J., & Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 46-48.

Duffield, G., & Grabosky, P. (2001). *The psychology of fraud. Trends and Issues in Crime and Criminal Justice,* 199, 1-6. Canberra, Australia: Australian Institute of Criminology.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1-23.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2009). Cotan evaluation system for the quality of tests (In Dutch). Amsterdam, Netherlands: NIP. urn:nbn:nl:ui:29-346184.

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology, 97*, 866–80.

Foster, D. (2009). Secure, online, high-stakes testing: Science fiction or business reality? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 31-34.

Gibby, R. E., Ispas, D., McCloy, R. A., & Biga, A. (2009). Moving beyond the challenges to make unproctored internet testing a reality. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 64-68.

Gino, F., & Ariely, D. (2012). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology, 102*, 445–459.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83*, 314-320.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty-Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.

Kaminsky, K. A., & Hemingway, M. A. (2009). To proctor or not to proctor? Balancing business needs with validity in online assessment. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 24-26.

Landers, R. N., & Sackett, P. R. (2012). Offsetting performance losses due to cheating in unproctored Internet-based testing by increasing the applicant pool. International *Journal of Selection and Assessment, 20*, 220-228.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255-275.

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology, 84*, 817-824.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., …Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Nye, C. D., Do, B. R., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment, 16*(2), 112-120.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and "paper & pencil" testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56*, 733-752.

Potosky, D., & Bobko, P. (2004). Selection testing via the Internet: Practical considerations and exploratory empirical findings. *Personnel Psychology, 57*, 1003-1034.

Ryan, A. M., Inceoglu, I., Bartram, D., Golubovich, J. G, Grand, J., Reeder, M….Yao, X. (2015). Trends in testing: Highlights of a global survey. In I. Nikolaou, & J. K. Oostrom (Eds), *Employee recruitment, selection,* *and assessment: Contemporary issues for theory and practice* (pp. 136-153). New York, NY: Psychology Press.

Scott, J. C., & Lezotte, D. V. (2012). Web-based assessments. In N. Schmitt (Ed). *The Oxford handbook of personnel assessment and selection* (pp. 485-513). New York, NY: OUP.

Shepherd, W. J., Do, B. R., & Drasgow, F. L. (2003, April). *Assessing equivalence of online noncognitive measures: Where research meets practice.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 2–10.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). *Unproctored Internet testing in employment settings. Personnel Psychology, 46*, 189-225.

RECEIVED 1/6/16 **ACCEPTED** 6/18/16