Industrial and Organizational Psychology, **2** (2009), 69–76.
Copyright © 2009 Society for Industrial and Organizational Psychology. 1754-9426/09

RESPONSE

Where Is the Unproctored Internet Testing Train Headed Now?

NANCY T. TIPPINS
Valtera Corporation

So, where is unproctored Internet testing (UIT) now? What has happened since 2006? Is the train stalled right outside the station, waiting on answers? Is it lurching forward, making slow but sure progress? Or, is it simply chugging in circles? The original focal article (Tippins, 2009) and the 13 subsequent commentaries are not comprehensive summaries of all the research and thought about UIT and can hardly be considered definitive as a collective body of work. In total, these articles probably do not convince many of the doubters. However, the articles collectively highlight the critical issues faced in UIT, suggest new ideas, and give guidance for the future, indicating the UIT train is at least moving. In this concluding article of the UIT series, I try to summarize where I believe we are in 2009 in terms of the important points that must be taken into consideration when contemplating UIT and emerging questions that industrial and organizational (I–O) psychologists must answer.

Important Points

Regardless of where one stands on UIT, there are some important points that need to be kept in mind. None of the following "important" points is new. Some have been apparent since UIT became a possibility. Others are new only to those of us who have not been paying much attention in the past few years. Nevertheless, each of them needs to be emphasized.

Cheating Occurs on Virtually All Tests and People of All Levels of Ability Cheat

The possibility of some cheating on employment tests seems likely. Although there is little if anything in the literature that indicates the extent of cheating on employment tests, several of the authors report indications of cheating, and several of the studies reported here note score differences between unproctored and proctored tests that may be interpreted as cheating. Burke (2009) reports data indicating that 45% of job applicants falsify work histories and would presumably be willing to cheat if given the opportunity. In addition, his Web patrols have been alarmingly successful at detecting sites that compromise the test materials of his company. Hense, Golden, and Burnett (2009) report an effect size of .32 between scores on a proctored and unproctored job simulation. Arthur, Glaze, Villado, and Taylor (2009) estimated that cheating, as defined by 1 *SEM* of the difference scores of a proctored test when compared to a parallel unproctored test, occurred, in as much as 7.7% of the sample when using a cognitive test that is speeded

Correspondence concerning this article should be addressed to Nancy T. Tippins. E-mail: ntippins@valtera.com

Address: Valtera Corporation, 301 North Main St., Suite 1103, Greenville, SC 29601

Nancy T. Tippins, Valtera Corporation.

https://doi.org/10.1111/j.1754-9434.2008.01111.x Published online by Cambridge University Press

to prevent cheating. The authors also note that cheating apparently occurs across the entire distribution of scores on both cognitive and noncognitive tests.

Cheating appears widespread, although the extent of cheating and the conditions that stimulate cheating are not well understood. An employer who uses any selection test, proctored or unproctored, may reasonably expect some cheating to occur and would be wise to consider the implications.

UIT Is Not a Single, Homogeneous Issue, and the Essential Question Is Not "Is UIT Good or Bad?"

Psychologists studying UIT have long noted that the acceptability of UIT varies according to many factors ranging from test content to the purpose of the testing. The acceptability of UIT used for development may be different from that of UIT used for selection; an unproctored biodata form may generate different concerns about cheating than an unproctored verbal reasoning test. Reynolds, Wasko, Sinar, Raymark, and Jones (2009) present a clear framework for decomposing UIT into all the conditions and processes that affect various outcomes.

If UIT is not a single, homogeneous issue, then one answer regarding its value that applies to all conditions is unlikely. In particular, the answer to the questions, "Is UIT good or bad?" and "Should I use UIT or not?" will not be universal; instead, the final evaluation will depend on the outcome desired, the test itself, the conditions under which it is administered, and test-taker perceptions. Similarly, the best practices around the use of UIT will vary according to the combination of factors involved.

The Evaluation of the Effectiveness of Any Testing Program Should Not Be Taken Out of the Context of the Entire Hiring Process

Hense et al. (2009) and Gibby, Ispas, McCloy, and Biga (2009) emphasize the importance of addressing the needs of the organization through the entire selection process. In practice, any testing program must be evaluated in the context of the entire hiring program beginning with the definition of the job requirements and the recruiting strategies and ending with employee selection and on boarding. As Hense et al. noted, UIT solved some important nontesting problems in the hiring process such as excessive time spent screening candidates and inconsistent execution of processes while undoubtedly creating or exacerbating the problems of cheating, candidate identity, and test security.

Validity Is Not the Only Issue of Concern

Several articles indicate that validity often decreases in unproctored settings when compared to proctored testing, although some studies show stable validities (Kaminski & Hemingway, 2009). The extant research suggests the validity of an unproctored test is lower than that of a proctored test but is often significantly greater than zero. All things equal, higher validities are good. However, if all things are not equal and the applicant pool increases substantially and the number selected remains constant, then the utility of the test increases: More highly qualified individuals can be identified. Similarly, the use of UIT may expand the applicant pool such that more members of protected classes are included. The user of any testing program, proctored or unproctored, must consider multiple factors (e.g., validity, adverse impact, selection ratio) when evaluating UIT.

Standards for Evaluating UIT Are the Same as Those for Any Other Test

The standards against which any test is judged remain the same regardless of the conditions under which it is administered. Concepts such as validity, reliability, and bias are still relevant and must be evaluated. Whether current guidelines need to be amended to explicitly state expectations regarding the standards for UIT as Beaty, Dawson, Fallaw, and Kantrowitz (2009) suggest is debatable; the expectation is not.

https://doi.org/10.1111/j.1754-9434.2008.01111.x Published online by Cambridge University Press

Proctored Testing Is Not the Gold Standard for Evaluating UIT

Bartram (2009) and Drasgow, Nye, Guo, and Tay (2009) point out that proctored testing is not immune from cheating and can hardly be presented as the gold standard for pristine testing practices. I–O psychologists who have managed large proctored testing programs are all too familiar with sleepy proctors, administrators who forget to account for testing materials, and candidate substitution.

Foster's description (2009) of technology-enhanced ways of monitoring tests reminds us that there are different ways to proctor test administration sessions and indicates there are new ways to monitor tests that do not require the physical presence of a proctor in the same room with a test taker. His descriptions also remind us that even technology-based solutions require some human intervention—someone must conscientiously monitor cameras or the data produced from the technology—to be effective. These humans too must remain attentive to their tasks.

A closely related concern is the test takers' perceptions of the test environment. One of the more surprising points made in this set of articles is Weiner and Morrison's (2009) finding that unproctored on-site testing received the worst ratings on factors such as noise compared to proctored on-site testing and unproctored remote testing. One might have expected the on-site unproctored venue to have fewer distractions in general when compared to those of remote unproctored sites. Clearly, there is no perfect standard against which to compare UIT.

New Questions

Although the research literature on UIT is growing, the list of questions is also expanding. Existing questions have not been answered conclusively, and new questions are emerging from recent research. The current set of articles leads me to add some new questions to those listed in my focal article (Tippins, 2009). (See Appendix.)

I have created two groups of questions. I have labeled one group "Questions with a Philosophical Answer." The answers to these will not be found in a research study; rather, the answers will be based on reason and must be determined by combining our profession's ethics, standards, and values as well as our own. The second group is labeled "Questions with a Research-Based Answer." Well-designed and executed research can inform our understanding of the issue and help answer these questions. These questions are briefly discussed below.

Questions With a Philosophical Answer

Are UIT tests different from a proctored test with the same validity? A fundamental question we must ask ourselves is what differences in tests matter. Are two tests with the same validities fundamentally different if one is given in a proctored setting and the other in an unproctored setting? Those of us who argue these tests are essentially the same will point out that in the long run both tests will have similar results at least within large groups. All things are rarely equal, so the psychologist must consider what happens if UIT increases the size, quality, and diversity of the applicant pool. Some of us will argue that the two tests are fundamentally different and point out the unproctored score is a less accurate indicator of an individual's true score on the construct if that person cheated on the unproctored test.

Many practitioners of I–O psychology make decisions about what kinds of tests to implement using nonpsychometric factors including costs, administrative feasibility, organizational expectations, and so forth. Interviews may be chosen over a cognitive ability test with a higher validity for many reasons. The decision to use UITs must be made on the basis of the same factors.

What is the acceptable balance between the rewards associated with unproctored testing and the risks? Almost no one would argue that unproctored testing is unflawed. Virtually all of us are aware of the problems of cheating, the difficulties of verifying applicant identity, concerns about

https://doi.org/10.1111/j.1754-9434.2008.01111.x Published online by Cambridge University Press

testing environments, and the consequences of compromised test security when using UIT. Yet the benefits are obvious too. The cost of testing can be greatly reduced, and the applicant pool is often widely expanded. So, when do the benefits outweigh the liabilities? If all unproctored testing does is transfer the site of testing from an employer's proctored premises to the applicant's location of convenience, there are still significant cost savings to the employer associated with shifting the equipment burden to the applicant and eliminating administration personnel. If, however, the unproctored testing expands the applicant pool by a factor of two or three or greater, and if that applicant pool contains more qualified candidates who are reluctant to take time from their jobs, and if that applicant pool is more diverse, are the benefits sufficient to outweigh the problems of unproctored testing?

The answers to such questions of course depend on the value one places on the benefits and the costs. Obviously, the benefits need to be worth more than the costs. Unfortunately, there seems to be no common metric for the value of cost reductions, expanded labor pools, or the opportunity for a cheater to succeed. Thus, the unresolved question is not just how much more benefits are worth when compared to the costs, but also how each advantage and disadvantage is measured.

What is the extent of a psychologist's obligations to ensure test standardization, test security, and applicant identity? The concepts of test proctoring, test security, and test standardization can be operationalized in many ways. For example, there can be no test proctoring at all, test takers may be asked to behave honorably in the test-taking situation, test takers may be observed via technology, or test takers may be directly observed by human proctors. Most of us agree that no form of proctoring is perfect; there are usually multiple ways to circumvent humans or technology. Similarly, the elements of a test and its administration may be standardized to varying degrees as can the level of test security provided. Consequently, we must ask how much effort to prevent cheating and protect test materials is sufficient and how much standardization is enough. What is the "reasonable effort" required by the Ethics Code? This question then begs the question of who defines what that effort should be and when enough is enough. Are I–O psychologists expected to do all that is humanly possible to eliminate cheating, protect test content, and standardize the testing experience? Or, can those efforts be tempered by costs and feasibility?

Should the burden of providing an appropriate testing environment be appropriately shifted to the test taker? For many of us, the line between the employer's responsibility and the test taker's is blurred. In traditional, proctored testing programs, the employer has assumed much of the responsibility for the testing environment but not all. Many testing programs admonish the test taker to eat breakfast and get a good night's sleep, clearly shifting the burden of personal fitness for testing to the individual. Other programs directly ask the question, "Is there any reason why you cannot do your best today?" and require the test taker to define what conditions are acceptable to him or her. In contrast, UIT shifts the entire burden of appropriate testing conditions to the individual. The individual must determine his or her personal fitness for taking the test as well as assess the appropriateness of the environment for test taking. Most of us are comfortable allowing the test taker to assume sole responsibility for personal fitness—who else could? But our comfort with assigning the responsibility for determining environmental acceptability gives us pause. Undoubtedly, some of us are uncomfortable with the shift from traditional testing practices. Others of us are aware that not everyone has the privilege of choosing his or her environment. A few of us believe the choices the test taker makes in this regard tell us something about the individual that may be job relevant.

Should our standard be absolute or relative? Is it better to have unproctored testing than random selection? Or, is it better to do no

https://doi.org/10.1111/j.1754-9434.2008.01111.x Published online by Cambridge University Press

testing unless it can be executed perfectly? Organizations such as the Starwood Hotels described in the Kaminski and Hemingway article (2009) seem to be faced with the prospect of testing in unproctored settings or no testing at all. Consequently, we must confront the question, is UIT testing with all its flaws better than no testing at all.

Questions With a Research-Based Answer

How does the amount of cheating that occurs in proctored testing compare to that which occurs in unproctored testing? Many authors lament the dearth of data on the amount of cheating that occurs in employment settings. There is virtually no information that compares the amount of cheating that occurs with various kinds of test content used for various purposes in settings with human proctors directly observing, settings with human proctors observing test performance remotely using technology, and settings with no proctoring. Without a baseline, it is difficult to evaluate the relative effectiveness of various kinds of interventions designed to prevent cheating. We do not know if the inattentiveness of human proctors who directly observe testing sessions is also characteristic of those proctors who observe banks of cameras trained on test takers. We do not know if the opportunity to cheat increases the cheating behavior or simply makes it easier to do so effectively.

How does test validity and test security change over time when UIT is used? Pearlman (2009) suggests that cheating and test taker authentication have variable effects to the extent that the validity of the test is not maintained over time. In contrast, Beaty et al. (2009) report several studies that show stable scores over time. An important research effort needs to focus on how validity changes over time when UIT is employed. If a test is validated using applicants in an unproctored condition, can we expect the validity to decrease? Do test security problems proliferate over time? The research to answer such questions can be difficult to execute, but it will be necessary.

What constitutes evidence of cheating? Bartram (2009) points out that positive evidence of cheating is needed rather than circumstantial evidence. No one who understands test scores or has managed a testing program is likely to accuse someone of cheating based solely on score differences. Mere differences in unproctored and proctored test scores do not necessarily constitute evidence of cheating. Yet the questions remain: "What constitutes this 'positive evidence'?" and "How much evidence is needed to conclude cheating has occurred?"

What personal characteristics are related to attitudes regarding the test environment? The original set of questions asked if there were moderators of cheating behavior such as ability. Weiner and Morrison's (2009) finding that noncognitive test scores were correlated with ratings of the environment raises new questions about how different people respond to different testing environments and different levels of proctoring.

What changes to test format (e.g., speeded tests, Foster Item3) may reduce the incidence of cheating? Two articles in this set, Arthur et al. (2009) and Foster (2009), highlight new ways to deter cheating. Arthur et al. use a speeded cognitive ability test in an unproctored setting to decrease the effects of cheating. Foster's article stimulates ideas about how item formats may reduce cheating. Of course, this questions the comparability of new item formats to old ones particularly in terms of the construct measured; however, that may be a moot point if comparability is not required.

What forms of UIT research can support small-scale testing programs? Another research area that is particularly deficient is solutions for smaller scale testing programs. Many of the measures to prevent cheating are based on techniques that require large samples. Gibby et al. (2009) acknowledge that the UIT tools used at Proctor & Gamble are beyond the reach of smaller companies. Unless the smaller employer relies on off-the-shelf tests, UIT and the accompanying cheating deterrents may not be feasible.

https://doi.org/10.1111/j.1754-9434.2008.01111.x Published online by Cambridge University Press

Conclusions

In the waning months of 2008, we have a growing set of questions about UIT to be answered and a growing literature examining the effects of UIT and suggesting best practices. We seem to know more today than we did in 2006 both in terms of what questions to ask and what is known about UIT. In response to my own final question, "Where do we go from here?", I have five suggestions.

Start Answering the Research Questions Already Posed

We have already defined many of the important questions that need to be answered. Researchers like Beaty et al. (2009) have defined their research agendas. Now, scientists and practitioners must execute careful research to provide answers and direction. UIT occurs every day. Those of us who are involved must find a way to contribute to our understanding of UIT.

Honestly Confront the Philosophical Questions

The philosophical questions require us to consider our standards, our values, our ethics, and our responsibilities to our clients and their applicants. Constructive debate can help us understand the issues and formulate answers to these questions.

Collect and Share Accurate Cost Information About All Aspects of the UIT Process

One glaring deficiency in our warehouse of information is data on costs. We do not know how much verification testing costs; we do not know how much the development of replacements for compromised items and research on them costs. Without facts, we may delude ourselves into believing that UIT is cheaper than proctored testing or that proctored testing is actually more cost effective in the long run.

Money is a significant factor in almost every testing program. To fully comprehend the costs of UIT and its benefits, we must identify all the activities required and the costs associated with them as well as determine what effects UIT has on the hiring process and the costs associated with them. For meaningful comparisons to be made, cost information about the steps in proctored testing must also be accumulated. Gibby et al. (2009) provide a glimpse of the costs, "To reproduce the UIT tools we developed would require a multi million dollar budget and access to tens of thousands of respondents—both candidates and employees." Much more, specific information is needed to fully evaluate the cost of UIT.

Explore Technological Means of Monitoring

Technology changes rapidly and is likely to provide new means of identifying candidates and preventing cheating. I–O psychologists must stay abreast of these technologies and the opportunities they present.

Attend to the Needs of Employers Who Lack Large Pools of Applicants and Significant Resources to Expend on Test Research

Except for the largest employers, most lack large pools of applicants or employees who can serve as subjects in research to detect and deter cheating. In today's economy, many employers lack the monetary and human resources to pursue such research. Our research needs to identify solutions to UIT that do not depend on large samples.

If we do not undertake these activities, someone will answer the generic UIT question for us. We may find ourselves in a very uncomfortable seat on a fast-moving UIT train.

References

Arthur, W. A. 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.

Bartram, D. (2009). The International Test Commission guidelines on computer-based and Internet-delivered testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 11–13.

Beaty, J. C., Dawson, C. R., Fallaw, S. S., & Kantrowitz, T. M. (2009). Recovering the scientist-practitioner

https://doi.org/10.1111/j.1754-9434.2008.01111.x Published online by Cambridge University Press

model: How I-Os should respond to unproctered Internet testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 58–63.

Burke, E. (2009). Preserving the integrity of online testing. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 35–38.

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.

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.

Hense, R., Golden, J., & Burnett, J. (2009). Making the case for unproctored Internet testing: Do the rewards outweigh the risks? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 20–23.

Jones, J. A. (2009). UIT or not UIT? That is not the only question. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 52–57.

Kaminski, 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.

Pearlman, K. (2009). Unproctored Internet testing: Practical, legal, and ethical concerns. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 14–19.

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.

Weiner, J. A. & Morrison, J. D., Jr. (2009). Unproctored online testing: Environmental conditions and validity. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*, 27–30.

Appendix
Revised Table of Open Questions Regarding UIT

Extent of cheating

1) What kinds of cheating (e.g., outside assistance, substitution of test takers) take place and what is the rate of each?
2) Does the rate of cheating vary by demographic subgroups?
3) Are there other moderators of cheating behavior (e.g., ability)?
4) How do the consequences of test results affect cheating behavior?
5) On what types of tests (e.g., cognitive tests) or test formats (e.g., multiple-choice tests) is cheating most likely to occur?
6) What deterrents to cheating (e.g., warnings, threats of verification testing, honor statements) are most effective?
   a. What changes to test format (e.g., speeded test, Foster Item3) may reduce cheating?
7) Does the kind of cheating or the rate of cheating change over the life span of a UIT program?
8) How does the amount of cheating that occurs in proctored testing compare to that which occurs in unproctored testing?
9) What constitutes evidence of cheating?

Impact on validity

10) What is the validity of inferences made from test results when the test was administered under unproctored conditions?
11) What is the acceptable tradeoff between decrements in validity and utility gains (e.g., cost savings associated with UIT, larger applicant pool)?
12) Are UIT tests different from a proctored test with the same validity?
13) How does test validity and test security change over time when UIT is used?

Impact on the organization

14) What effect does UIT have on each step of the hiring process?
   a. Does UIT affect the number and quality of applicants and test takers?
   b. What features of UIT (e.g., speed of hiring, accessibility of testing process) affect the number and quality of applicants and test takers?
   c. Does UIT affect the number of qualified applicants who progress to the next stages of hiring (e.g., interviews, background checks, drug tests)?

https://doi.org/10.1111/j.1754-9434.2008.01111.x Published online by Cambridge University Press

15) What effect does UIT have on the organization in terms of the capabilities of the new hires and their integrity?
   a. Is there a measurable decrease in abilities of new hires due to some percentage of the applicants hired having cheated and subsequently being hired?
   b. Is there a measurable difference in performance between those who test in a proctored setting and those who test in an unproctored setting?
   c. Are more questions about individual integrity raised in groups of people who took tests in an unproctored setting compared to those who took tests in a proctored setting?
16) What are the costs and benefits to UIT and the potential solutions?
   a. Is electronic surveillance cost effective?
   b. Is verification testing cost effective? How much do equivalent tests cost to develop? What is the cost of large item banks? How often must items be replaced?
17) Should our standard be absolute or relative? Is it better to have unproctored testing than random selection? Or, is it better to do no testing unless it can be executed perfectly?
18) What forms of UIT research can support small-scale testing programs?

Impact on the individual

19) How does UIT affect test performance? Does an individual do as well in an unproctored setting as he or she would in a proctored setting?
20) How does UIT affect individual applicant attitudes toward an employer? What are incumbents' opinions regarding UIT?
21) Should the burden of providing an appropriate testing environment be appropriately shifted to the test taker?
22) What personal characteristics are related to attitudes regarding the test environment?

Ethics and legal issues

23) Are there ethical concerns about UIT?
24) What is the acceptable balance between the rewards associated with proctored testing and the risk?
25) What is the extent of a psychologist's obligations to ensure test standardization, test security, and applicant identity?
26) Do the ethics of UIT depend on the type of test item? For example, would administration of a cognitive ability test be unethical in an unproctored setting while administration of biodata would not?
27) What are likely to be the legal challenges to UIT in the United States and other parts of the world?

Guidelines

28) How effective are the current guidelines?
29) What policies or guidelines need to be implemented?
30) What should SIOP/APA/EAWOP be doing?
31) Do the current ITC Guidelines represent good practice or should it be amended?
   a. If it should be amended, in the light of more recent research and experience, what changes should be made?

Technology

32) How can technology effectively monitor behavior in testing sessions?
33) Is electronic monitoring as effective in candidate identification and cheating detection as human proctoring?

https://doi.org/10.1111/j.1754-9434.2008.01111.x Published online by Cambridge University Press