Industrial and Organizational Psychology, **2** (2009), 14–19.
Copyright © 2009 Society for Industrial and Organizational Psychology. 1754-9426/09

Unproctored Internet Testing: Practical, Legal, and Ethical Concerns

KENNETH PEARLMAN

The thesis of this brief comment is that unproctored Internet testing (UIT), when used to make important decisions about individuals, is subject to serious ethical and legal challenges, as well as a number of organizationally harmful consequences that may outweigh any real or perceived benefits. This situation is both alarming and troubling, as the deployment of UIT has become widespread in the United States in recent years, to the point where it is an almost standard part of the application process in some economic sectors. For example, the UIT-based administration of some type of personality test immediately following completion of an online employment application—with no attempt to ensure test security, authenticate applicant identity, or control (or even discourage) cheating and with no follow-up (or "verification") testing (as discussed below)—is now probably the modal application protocol for nonexempt jobs in nationwide retail chains. Although definitive usage data are lacking, it is quite possible, even probable, that UIT accounts for the majority of individual employment test administrations that currently take place in the U.S. private sector. In conjunction with the ethical and legal challenges just mentioned, on which I elaborate below, this means there is a strong likelihood that most contemporary U.S. employment testing practice is in violation of both ethical standards of psychological practice and accepted professional testing standards.

This discussion is limited to UIT settings, that is, situations where there is no (either direct or technology enabled) human or electronic monitoring of individuals while they are taking a test. When discussing such situations, the most critical distinction is that between low-stakes testing applications (LSTAs) and high-stakes testing applications (HSTAs). LSTAs are situations in which testing consequences or outcomes are of relevance or value only to the individual tested and have limited or no direct effects on or consequences for other individuals or institutions (e.g., the provision of sample or practice tests or self-diagnostic tests). HSTAs are situations in which test results are used to make evaluations, judgments, or decisions that are highly consequential to an individual and also have consequences or outcomes beyond the individual tested (e.g., as in the cases of employee selection tests or professional credentialing tests).

In contrast to the LSTA context, where UIT can be highly beneficial to both individuals and organizations, as I have described elsewhere (Tippins et al., 2006), UIT is never appropriate for HSTAs, for two major reasons. First, the fundamental, widely acknowledged (cf. Bartram, 2006; Naglieri et al., 2003), and interrelated problems of candidate cheating, candidate identity authentication, and test and item security are of significant concern only in the HSTA context. (For one thing, these problems are much likelier to arise in HSTAs than LSTAs,

Correspondence concerning this article should be addressed to Kenneth Pearlman. E-mail: kenpearlman@comcast.net

Address: Creative Personnel Management Consulting, 7848 Camminare Drive, Sarasota, FL 34238

https://doi.org/10.1111/j.1754-9434.2008.01099.x Published online by Cambridge University Press

and for another, issues of cheating and test security with respect to publicly available or self-diagnostic tests are of no broader institutional or societal concern.) Second, these problems can be mitigated to varying degrees via process- and technology-related, and even statistical, interventions, but they can never be entirely eliminated.[1]

There are three types of consequences to these ultimately intractable problems—practical, legal, and ethical. More specifically, UIT, as typically deployed in HSTA contexts, (a) will likely fail to benefit, if not actually harm, its sponsoring organizations in both direct and indirect ways; (b) almost inevitably violates professional testing standards, which makes it problematic to defend legally if challenged; and (c) violates many facets of the ethical standards for psychologists. I briefly elaborate on the most salient aspects of each of these issues below.

Direct and Indirect Practical Consequences of UIT

The most obvious direct consequence of the interrelated problems of cheating, candidate authentication, and test security—to which all high-stakes applications of UIT are subject—is that it results in test scores that may not accurately reflect an individual's standing on whatever the test is measuring, thereby reducing its validity. There is ultimately no way to overcome these security- and cheating-related problems in unproctored settings, other than through technologies (e.g., video cameras or other means of remote surveillance) that would mean the setting is no longer unproctored, it is just proctored remotely.

The most commonly proposed or applied remedies for these problems are either very limited or misconceived. Perhaps the most common one is the "test–retest" approach, wherein an unproctored test is used as a prescreen, followed by proctored administration of the same test as a "verification" or "confirmation" test. Variations of this approach are often deployed, such as using a computer-adaptive version of the unproctored screening test in an attempt to minimize some forms of cheating (those based on test or item compromise); such methods are of course irrelevant to the probably far more common form of cheating in UIT settings—test takers being assisted (or replaced) by others. Setting aside the thorny ethical and legal dilemmas of what to do about applicants suspected of cheating based on a drop in their scores between the Internet and the verification test, the practical problem with such approaches is that they largely undermine the purpose and attractiveness to organizations of implementing UIT in the first place, which is to have a quick, seamless, largely automated, and "candidate friendly" application—selection process that will reduce costs by eliminating on-site or proctored testing. By contrast, the test–retest approach (and its variations) adds steps (and hence, time and cost) back into the process, as well as constituting a tacit admission by the organization that UIT results cannot be relied upon, which may have negative indirect consequences, as discussed below.

There are other potential problems with these sorts of strategies. For example, if the prescreen cut score had been derived from unproctored test administrations, it could have been artificially inflated because of the effects of cheating. This would result in inappropriate disqualification of some proportion of qualified candidates (i.e., false negatives) before they ever get to the verification test, which might succeed in revealing some cheaters (although even this is by no means straightforward) but cannot recoup the lost "honest"—and qualified—candidates. Even if the prescreen cut score had not been inflated (e.g., if derived from

1. I fully recognize that, in addition to the LSTA–HSTA distinction, there are other distinctions—such as between ability and personality measures and between speeded and power tests—that at least some research suggests may moderate the precise frequency and dynamics of these problems (e.g., personality and ability tests are likely to be differentially vulnerable to different cheating strategies). My view is that such findings, even if verified, still do not call into question the inevitable manifestation of such problems in all high-stakes UIT contexts. For this reason, along with space constraints, I do not address such factors in this discussion.

https://doi.org/10.1111/j.1754-9434.2008.01099.x Published online by Cambridge University Press

proctored administrations), to the degree that cheating lowers test validity, the proportion of false negatives will still increase. In either case, the fundamental purpose of testing is being undermined, in that the use of a UIT prescreen results in more false negatives (i.e., the loss of more qualified applicants) than would otherwise have occurred. More generally, as I have argued elsewhere (Tippins et al., 2006), to the extent that the validity of a test has been undermined because of cheating and other effects of nonstandardized administration (a consequence that, in high-stakes applications, is likely exacerbated over time as a result of the cumulative effects of test compromise and organized cheating), UIT's supposed advantage in reducing the size of the applicant pool and realizing subsequent hiring process cost or time savings is neutralized because this step is not passing along *better* (more qualified) people to successive selection process steps, only *fewer* people, an outcome that could be accomplished in any number of other simpler and less expensive ways.

In addition to these direct consequences, there are likely to be undesirable indirect or spillover effects of UIT. For example, there is mounting evidence from several lines of research (e.g., Schneider's attraction—selection—attrition framework and the organizational justice literature) that (a) assessment methods provide applicants with information that they use to evaluate their congruence with the culture and values of an organization, make job acceptance decisions, and set expectations regarding acceptable behavior within the organization (Anderson & Ostroff, 1997); and (b) candidate perceptions of the procedural fairness of an organization's selection tools are related to perceptions of organizational attractiveness, likelihood of speaking favorably about the organization to others, and likelihood of accepting a job offer (Hausknecht, Day, & Thomas, 2004). A UIT-based selection tool that is so readily and obviously "gameable" and subject to compromise, and hence lacking in integrity, is likely to have negative ramifications on such perceptions and outcomes. (Some might argue that this is not substantively different from other aspects of the application process that are subject to applicant "cheating," in the form of self-misrepresentation, such as in stated employment history or employment interview responses. I would contend that one of many differences that such a comparison overlooks is the complete absence, in the typical UIT context, of a good-faith effort to ensure test standardization, test security, and applicant identity, which are basic prerequisites for meeting ethical and professional standards of test use, as I discuss later.) It is not unreasonable to speculate that, at a broader level, such perceptions would inevitably be extended to the more general applicant pool (thereby inhibiting an organization's recruiting ability) and to the public at large, thereby harming an organization's reputation among its customers and business partners.

Another indirect effect of UIT is that it arguably promotes the production and proliferation of bad (i.e., poorly developed or unvalidated) tests, as noted by the American Psychological Association's Task Force on Psychological Testing on the Internet, which warned of "the easy entry of many nonprofessionals selling tests that do not meet professional standards" (Naglieri et al., 2003, p. 33).

Legal Consequences of UIT

One potential legal jeopardy is associated with UIT centers around issues of either differential access of protected groups to testing equipment or facilities—an issue that is likely to decrease in importance over time because of the increasing prevalence of Internet access—or differential test performance (including differential cheating rates) in such settings—an issue about which there is presently virtually no relevant research data. However, a second type of legal jeopardy arises should an organization have to defend its UIT practices. The reduced test validity likely to result from cheating or test security breaches is not a legal jeopardy per se because under the *Uniform Guidelines* (Guidelines; Equal Employment Opportunity

https://doi.org/10.1111/j.1754-9434.2008.01099.x Published online by Cambridge University Press

Commission, Civil Service Commission, Department of Labor, 1978), it is perfectly legal to use a bad or compromised test, as long as it has no "adverse impact" as the Guidelines define it. However, in cases in which adverse impact is found, the guidelines would require an organization to demonstrate the validity of its UIT practices. This is where things become problematic.

The guidelines sets forth standards for demonstrating test validity, while also recognizing—and striving for consistency with—existing professional testing standards, the most important of which are the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education, 1999) and the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 2003). The Guidelines and both sets of professional standards are unified in requiring standardized test administration as a prerequisite for both evaluating test validity (i.e., conducting validation studies) and maintaining or ensuring test validity in ongoing operational testing programs. Standardized test administration, as defined in the AERA et al. (1999) *Standards*, means "maintaining a constant testing environment and conducting the test according to detailed rules and specifications, so that testing conditions are the same for all test takers" (p. 182). As elaborated in specific standards related to test administration, this implies such things as (a) a testing environment that furnishes "reasonable comfort with minimal distraction" (Standard 5.4); (b) making "reasonable efforts" to "assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means" (Standard 5.6); and (c) "protecting the security of test materials at all times" (Standard 5.7). Further, the comment section of Standard 5.6 states that for "large-scale testing programs where the results may be viewed as having important consequences" (i.e., HSTAs), this standard calls for "stipulating requirements for [test taker] identification," "providing continuous monitoring of the testing process," and designing "procedures to minimize the possibility of cheating" (AERA et al., 1999, p. 64).

It is readily apparent that UIT, by definition, violates both the letter and the spirit of such standards. [It is perhaps also worth noting in passing that the lack of standardization inherent in UIT calls into question whether a measure administered in such a fashion can even legitimately be considered a "test," in light of that term's common definition in both the AERA et al. *Standards* and the SIOP *Principles* as "a measure or procedure in which a sample of an examinee's behavior in a specified domain is obtained, evaluated, and scored using a standardized process" (SIOP, 2003, p. 71, emphasis added). Most significantly, given that a test's validity can be neither accurately estimated nor maintained over time in the face of the unknown (and, arguably, unknowable) and highly variable effects of cheating, test compromise, and candidate authentication problems, a credible legal defense of the validity of any UIT measure would appear problematic. Another facet of this problem lies in the guidelines' (and both sets of professional standards') requirement that the conditions of a test's actual use replicate as closely as possible the standardized conditions under which its validation research was conducted. This means that tests validated in a proctored setting but deployed in a UIT context (which probably reflects dominant current practice, although there are no reliable data on this) would be similarly problematic to defend.

Ethical Consequences of UIT

The use of UIT entails likely violation of multiple sections of the "assessment" portion (Section 9) of APA's *Ethical Principles of Psychologists and Code Of Conduct* (APA, 2002), which SIOP endorses. Several examples of this, with brief explanations, are provided below.

> (a) Psychologists administer, adapt, score, interpret, or use assessment techniques, interviews, tests, or instruments in a manner

https://doi.org/10.1111/j.1754-9434.2008.01099.x Published online by Cambridge University Press

> and for purposes that are appropriate in light of the research on or evidence of the usefulness and proper application of the techniques
>
> (b) Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested. When such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation. (Section 9.02, Use of Assessments)

Tests that have not been validated under the UIT conditions in which they are actually deployed would be in violation of both parts of this principle. Anecdotal evidence suggests this practice is widespread. Part b would additionally be violated to the degree that psychologists do not disclose (e.g., in the test's manual or technical report) the inapplicability of validation results obtained in a standardized (i.e., non-UIT) context to a test's actual use in a UIT context.

> Psychologists who develop tests and other assessment techniques use appropriate psychometric procedures and current scientific or professional knowledge for test design, standardization, validation, reduction or elimination of bias, and recommendations for use. (Section 9.05, Test Construction)

This principle is compromised by, among other things, the violation of all accepted professional and legal standards for standardized test administration inherent in the UIT context, as described in the previous section.

> Psychologists who offer assessment or scoring services to other professionals accurately describe the purpose, norms, validity, reliability, and applications of the procedures and any special qualifications applicable to their use. (Section 9.09a, Test Scoring and Interpretation Services)

Under this principle, psychologists who provide or sell UIT products or services for high-stakes applications would have to disclose, as a "special qualification" applicable to the use of such products or services, the fact that they are highly subject to various forms of abuse and compromise, and would also have to disclose that any data (e.g., normative, validity, reliability) reported or available on such products or services (e.g., in the test's manual or technical report) that had not been gathered under UIT conditions is meaningless. Such disclosure is probably very rare, placing most UIT vendors in violation of this principle.

> Psychologists make reasonable efforts to maintain the integrity and security of test materials and other assessment techniques consistent with law and contractual obligations, and in a manner that permits adherence to this Ethics Code. (Section 9.11, Maintaining Test Security)

The public nature of the Internet, combined with the absence of proctoring, fundamentally undermines any meaningful degree of test material integrity and security in the UIT context. Such organizational countermeasures as computer-adaptive tests or randomized item presentation can mitigate but not eliminate this problem.

Naglieri et al. (2003) provide an in-depth discussion of the serious ethical challenges raised by typical UIT practices as noted above, as well as in many other sections of APA's *Ethical Principles*.

Conclusion

As I have stated previously (Tippins et al., 2006), the practical, legal, and ethical problems with UIT outlined here lead me to conclude that high-stakes applications of UIT are misguided and inappropriate *in principle* because the integrity of the testing process, and consequently the results it produces can never be assured to the "reasonable" degree required by professional standards. As a result, psychologists who engage in, promote, or profit from such applications create both ethical and legal risks for themselves and their organizations,

https://doi.org/10.1111/j.1754-9434.2008.01099.x Published online by Cambridge University Press

while very likely enabling various other organizationally detrimental consequences.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*, 1060–1073.

Anderson, N., & Ostroff, C. (1997). Selection as socialization. In N. Anderson & P Herriot. (Eds.), *International handbook of selection and assessment* (pp. 413–440). New York: Wiley.

Bartram, D. (2006). Testing on the internet: Issues, challenges and opportunities in the field of occupational assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances* (pp. 13–37). London: Wiley.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*, 38295–38309.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639–683.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., et al. (2003). *Psychological testing on the internet: New problems, old issues* (Report of the Task Force on Psychological Testing on the Internet). Washington, DC: American Psychological Association.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., et al. (2006). Unproctored internet testing in employment settings. *Personnel Psychology, 59*, 189–225.

https://doi.org/10.1111/j.1754-9434.2008.01099.x Published online by Cambridge University Press