Industrial and Organizational Psychology, **2** (2009), 35–38.
Copyright © 2009 Society for Industrial and Organizational Psychology. 1754-9426/09

Preserving the Integrity of Online Testing

EUGENE BURKE
SHL Group Ltd.

A key issue that has emerged is the security of Internet testing (Tippins et al., 2006). This can be readily understood in the context of statistics such as those offered by the background checking company Automatic Data Processing Inc. (2008), which reports that 45% of U.S. job applicants falsify work histories. To that can be added popular books such as *Freakonomics* (Levitt & Dubner, 2005), which asks the question in the second chapter: What do schoolteachers and sumo wrestlers have in common? The reader finds the answer is cheating. So, in an era of rapid change, new technologies, uncertainty, and concerns about trust, one can understand that the security of Internet testing is a legitimate concern.

But, as Tippins (2008) stated at a recent Society for Industrial and Organizational Psychology symposium on online testing, "The Internet testing train has left the station!" This reinforces the message that one of the key challenges to our science and practice is how we can effectively and practically defend the security and fairness of Internet testing. I will offer a few thoughts based on my work with colleagues on developing Internet testing solutions for clients in Asia, North America, and Europe (including the UK), experience gained from over 1.6 million ability tests delivered online in 19 languages over the past 7 years (Burke, 2008).

Correspondence concerning this article should be addressed to Eugene Burke. E-mail: eugene.burke@shlgroup.com

Address: Director of Science and Innovation, SHL Group Ltd., The Pavilion, 1 Atwell Place, Thames Ditton, Surrey KT7 0NE, United Kingdom.

Eugene Burke, SHL Group Ltd.

Managing Security Threats

First, let me make the assumption that the technology used to support Internet testing is itself secure in protecting the data that it manages and captures. That is, the information technology architecture and infrastructure are secure and robust attempts to access test content, scoring keys, and test score data. The focus on security then moves to the content of the tests themselves as revealed through their deployment online, and, most importantly, to the scoring keys used to turn responses to the test items into scores that will be used in decision making. The key risk is whether that content can be fraudulently accessed, which inevitably leads us to the issues of cheating and the actions of pirates in facilitating cheating on a test. It also leads us to focus on ability tests or any test for which there is only one correct answer to each item or where an instrument contains discrete answer patterns that can be associated with a higher score and where those patterns can be discovered and can be either memorized, recognized, or learned to result in a falsely inflated score for a candidate.

Procedures developed to defend against fraudulent access have to operate at two levels. The first is within the live assessment itself and includes clear rules for the test

https://doi.org/10.1111/j.1754-9434.2008.01104.x Published online by Cambridge University Press

taker about what is expected of and what is appropriate (and what is not) by designing the test to preserve the security of the scoring key and by incorporating in the overall testing or assessment process checks on the validity of scores and the authenticity of candidates, all of which act as a strong incentive to subscribe to honest test taking. Let me offer two quotes to support this first level of test security, the first of which emphasizes the need to set clear ground rules and is taken from Cizek (1999) and to wit (and to paraphrase): cheating is "deprive[ing] of something valuable by the use of deceit or fraud. . . . In testing specifically, cheating is violating the rules." Honesty contracts are familiar instruments to our field as used with assessments such as biographical data questionnaires, and, within a framework such as that provided by the *Guidelines on the Rights and Responsibilities of Test Takers* (American Psychological Association, 1996), do offer a clear basis for communicating the "rules" as Cizek refers to them.

However, as recent research in social psychology indicates (e.g., Vargas, von Hippel, & Petty, 2004; von Hippel, Lakin, & Shakarchi, 2005), people will often rationalize what would be generally construed as inappropriate behaviors, and therefore, additional incentives to follow the rules will strengthen security. Stronger measures are encouraged by the second quote I offer in relation to security measures internal to the assessment, and it is taken from Tippins et al. (2006): "Any Internet test that administers the same set of items to all examinees is asking to be compromised. At the very least, the items should be administered in a randomized order. It would be better yet to sample items from a reasonably large item pool." The Internet tests that we deploy in our systems use a randomized testing model through which equivalent but different tests are constructed from Item Response Theory (IRT) calibrated item pools and that are delivered client side (via the test taker's PC) via encrypted software that is returned server side with item responses at the end of the testing session. The software used client side does not contain any scoring key information, which is held securely server side. Each test has a unique scoring key.

To this requirement for the design of Internet tests, we have added verification of scores in the Internet testing models that colleagues and I have developed (Burke, van Someren, & Tatham, 2006; Burke & Wright, 2008). These verification tests are short tests administered in a proctored setting and at a later stage when authentication of the test taker can be undertaken, and these verification tests provide data that is used to check the consistency of scores from the first stage of testing. I should emphasize that this is not a second or additional score and that the score of record in our systems is the first score as this has been used to effect decisions on the full pool of test takers. The data from the verification test operate much like a faking good check would on personality instruments. These verification tests serve to provide yet another defense of the validity of the first Internet test and can be deployed at any point in the process such as the last or penultimate stage in which the candidate pool has been reduced to the final short list. This requirement was established through client interviews conducted in the Asia Pacific region, North America, Europe, and the UK. These clients asked for the score of record to be the first test and that this test be of sound psychometric quality to make accurate and valid decisions so that it could then be carried forward throughout remaining stages of any process, and that the verification tests be long enough to provide an accurate check on the first score but short enough to be practical in terms of including in later assessments such as assessment centers and/or interviews.

Thus far, I have mentioned security features operating within the testing process itself. The second level of security involves frequent external checks of the process to determine if procedures to manage test security are actually working. In the case of the systems that colleagues and I have developed, we deploy what are referred to as Web patrols to search the Internet for sites for potential breaches of security such

https://doi.org/10.1111/j.1754-9434.2008.01104.x Published online by Cambridge University Press

as pirate sites, brain dump sites, or forums through which access to content is being inappropriately offered. The reason that such searches work is quite simple: pirates and colluders can only achieve their goals if they are able to transact with others, and the medium that best facilitates such transactions is the Internet. So, ironically, the medium for which security measures are deemed important is also a key medium that enables those measures to be monitored for effectiveness.

In our systems, Web patrols are supported by data forensic audits for which space in this article prohibits a detailed explanation, and the reader is directed to http://caveon.com/df_blog/, which is maintained by Dennis Maynes, Chief Scientist at Caveon, and to Burke (2006) where readers will find a reproduction of an article by Dennis that provides an executive summary of data forensics. Data forensic algorithms look for evidence of aberrant scores or scores that, from knowledge of the psychometric properties of the test and how the items and test scores should behave under honest test taking conditions, are highly unlikely. One example would be a pattern of fast response times coupled with high correct answer rates to items that may suggest someone has had access to the scoring key. Another example would be long response latencies to items coupled with few right answers or all wrong answers that may suggest that the test taker is not a real candidate but someone harvesting the test content for possible piracy or collusion later. A third example would be algorithms that look for matches in correct and incorrect answer strings to tests that would suggest collusion among test takers.

Taken together, these measures operate to maintain the integrity of Internet tests, they enforce test security through actively managing breaches and monitoring test taker behaviors by identifying potential test fraud through Web patrols and critical incident procedures for dealing with security breaches and through data forensics audits to detect levels of item exposure, as well as designing cheat resistance into the testing procedures through equivalent randomized tests deployed via the Internet and proctored verification tests at later stages of the process.

Are These Measures Necessary and Do They Work?

In the past 3 years as we have evolved the security measures supporting our Internet testing systems, we have conducted Web patrols in English and Chinese (and more recently have begun to conduct patrols in French and Dutch with a rolling language program). Through these patrols, we have identified 30 sites over the past 18 months that have been qualified at three levels of risk using our critical incident procedures. Of those sites detected, 18 have been classified as high with 4 of those being sites in England and 14 of those sites operating in China (i.e., 40% of sites detected have been classified as high risk). Most of these sites, including ones on eBay, have been removed through simple letters or through invoking policies through the site operator or the Internet provider, but the 18 detected were offering content that would constitute a clear threat if the test had been a fixed set of items. These Web patrols have also shown that test security is an international issue with Chinese sites offering content in English. This is not surprising when one considers that many companies headquartered in Europe or North America also operate in China and have imported testing procedures from their home countries. We have conducted regular cycles of data forensics checks on tests and content threatened by such breaches, and the data show that the randomized testing model very clearly defends against collusion. These data forensics checks have also shown that there has been minimal impact from the exposure of item content. This can be seen as a result of the test design principles used because the items exposed do not have their scoring key available with them at the time of testing, and because items are configured in different permutations in the randomized testing model and are therefore unlikely to appear together in any one test.

I include here a case that I hope conveys the positive impact of the security measures

https://doi.org/10.1111/j.1754-9434.2008.01104.x Published online by Cambridge University Press

that I have described on the sense of fair play among test takers. The following is a verbatim extract from a UK site tracked through Web patrols on January 20, 2007 (all times are GMT). **Posted at 12:09** I got 20 real SHL numerical test questions. If someone needed, send me e-mail <e-mail address given>. Only 20 pounds. If you want to pay me a little money, then push you to the assessment stage, just e-mail me. Worth or not, you decide. **Posted at 16:01** There are a number of reasons why everyone should ignore the original poster. Firstly I have taken a number of SHL numerical tests and by and large the questions are of a similar level but different. Yes, there may be some that are repeated but out of 20 questions there is not much. More saliently, lets assume that by some miracle all 20 questions come up. Yes, you will pass but will be retested at the next round . . . so a waste of money and time. I really hate it when cretins like <Web name cited> take advantage of people. If she really wanted to help, she would have offered them for free. You have been warned . . . **Posted at 17:23** I agree, what a stupid thing to advertise. 1. It is not probable that you will get these 20 questions when you take the test (this person then goes on to list another nine reasons why the first posting is unhelpful becoming increasingly emotional about the posting and the person who posted it, hence the termination of the account of this thread at this point).

In conclusion, I do not claim that the procedures we have developed represent the only solution to the security of Internet testing or that our work is by any means completed. New challenges require new thinking that will continue to evolve. Indeed, these new developments do often show how we can strengthen the security and fairness of testing whether testing is proctored or not and have frequently shown us how insecure traditional proctored testing, the supposed benchmark for secure testing, is.

References

American Psychological Association (1996). *The rights and responsibilities of test takers: Guidelines and expectations*. Washington. DC: Author. Retrieved on July, 28, 2008 from www.apa.org/science/ttrr.html

Automatic Data Processing Inc. (2008). *2008 ADP screening index*. Retrieved on July, 28, 2008 from www.adp.com/media/press-releases/2008-news-releases/adp-annual-pre-employment-screening-index.aspx

Burke, E. (2006). *Better practice for online assessment*. Thames Ditton, UK: SHL. Retrieved on July 28, 2008 from www.shl.com/SHL/en-int/Thought_Leadership/White_Papers/White-Papers.aspx

Burke, E. (2008, April). Preserving the integrity of online testing. In N. T. Tippins (Chair), *Internet testing: Current issues, research solutions, guidelines, and concerns*. Symposium conducted at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Burke, E., van Someren, G., & Tatham, N. (2006). *Verify range of ability tests: Technical manual*. Thames Ditton, UK: SHL. Retrieved on July, 28, 2008 from www.shl.com/SHL/en-int/Products/Access_Ability/Access_Ability_List/verify.aspx

Burke, E., & Wright, D. (2008, January). *Defending the validity of online tests in an era of cheating and piracy*. Paper presented at the Annual Conference of the Division of Occupational Psychology of the British Psychological Society, Stratford-upon-Avon, UK.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.

Levitt, S. D., & Dubner, S. J. (2005). *Freakonomics*. New York: Penguin.

Tippins, N. T. (2008, April). *Internet testing: Current issues, research solutions, guidelines, and concerns*. Symposium presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

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.

Vargas, P. T., von Hippel, W., & Petty, R. E. (2004). Using partially structured measures to enhance the attitude-behavior relationship. *Personality and Social Psychology Bulletin*, *30*, 97–211.

von Hippel, W., Lakin, J. L., & Shakarchi, R. J. (2005). Individual differences in motivated social cognition: The case of self-serving information processing. *Personality and Social Psychology Bulletin*, *31*, 1347–1357.

https://doi.org/10.1111/j.1754-9434.2008.01104.x Published online by Cambridge University Press