CLEAR Exam Review

Volume XXII, Number 1
Spring 2011

A Journal

CLEAR Exam Review is a journal, published twice a year, reviewing issues affecting testing and credentialing. CER is published by the Council on Licensure, Enforcement, and Regulation, 403 Marquis Ave., Suite 200, Lexington, KY 40502.

Editing and composition of this journal have been written by Prometric, which specializes in the design, development, and full-service operation of high-quality licensing, certification and other adult examination programs.

Subscriptions to CER are sent free of charge to all CLEAR members and are available for $30 per year to others. Contact Stephanie Thompson at (859) 269-1802, or at her e-mail address, sthompson@clearhq.org, for membership and subscription information.

Advertisements and Classified (e.g., position vacancies) for CER may be reserved by contacting Janet Horne at the address or phone number noted above. Ads are limited in size to 1/4 or 1/2 page, and cost $100 or $200, respectively, per issue.

Coeditor
Michael Rosenfeld, Ph.D.
Educational Testing Service
Princeton, NJ 08541-0001
mrosenfeld@ets.org

Coeditor
F. Jay Breyer, Ph.D.
Educational Testing Service
Rosedale Road, MS13-R
Princeton, NJ 80541-0001
fbreyer@ets.org

CLEAR Exam Review

VOLUME XXII, NUMBER 1 SPRING 2011

Contents

 ISSN 1076-8025

The Importance of Data Forensic Applications in Today's Testing Programs

JERRY L. GORHAM, Ph.D.[1]
Pearson, Bloomington, MN

ADA WOO, Ph.D.
National Council of State Boards of Nursing, Chicago, IL

As the world of testing has grown larger, more technology-driven, and increasingly global, concerns over test security, standardized conditions, and other threats to the valid interpretation of examination results have become more critical. Innovation and growth often require new attempts to answer old questions such as: "Has this assessment been compromised?", "How do we identify likely cheaters?", or "What is the risk that our test items could be stolen?"

Data forensics has become an important set of tools for measurement experts attempting to answer such questions. In the testing industry, *data forensics* refers to the careful investigation of measurement events for the purposes of classification, identification, and interpretation of rare or unusual incidents. Forensics also assists in establishing the normal conditions under which standardized test administrations occur so as to highlight any unusual conditions that may require further investigation or policy changes. Data forensics generally produces likelihood statements that assist stakeholders and sponsors in developing or modifying future policy and regulations for their exams.

There are a number of ways in which data forensics can be useful to a given testing program. ***Security risk management*** is a principal application for data forensic analyses. This usually refers to methods, analyses, and processes that can detect various types of cheating behavior of examinees or others and their adverse effects on a testing program. The goal is to identify potential misbehavior as quickly as possible and to take remedial action to ensure that cheaters' scores are invalidated and the behavior can be disrupted or at least discouraged and minimized in the future. This category of data forensics is similar to criminal investigative forensics, which is known for sifting through evidence to find relevant pieces of information, then assembling those pieces into a story that forms a hypothesized picture of patterns of behavior. This comparison is even more appropriate when the connection between cheating and forensics is made. Aberrant response detection methods, collusion indexing analyses, and large retake discrepancy analyses are common types of security risk management forensics.

[1] Jerry Gorham is Manager of Measurement and Research at Pearson; Ada Woo is Senior Psychometrician at National Council of State Boards of Nursing. The authors gratefully acknowledge Mark Poole, Kirk Becker, Kathleen Gialluca, Sarah Hagge, and Steven Talbot for their assistance in preparation of this manuscript.

Psychometric consistency monitoring is another important application of forensic methods. Psychometric consistency monitoring refers to methods that detect deterioration in the quality of an examination but are not necessarily associated with deliberate cheating behaviors. These methods relate more generally to the tendency of examinations to undergo aging and degradation in their psychometric properties from repeated use over time and place. Curricular changes, population changes, overexposure of items, or other legitimate factors may render a particular test form less useful over time. An example might be a testing program that uses only one form of an examination repeatedly over a period of years. Although individual items may not be stolen, the testing population has become too familiar with the general attributes of the content and format so that the particular test form becomes easier and scores are inflated compared to their predecessors. Psychometric monitoring methods tend to focus on the global properties of the exam, its item or subscore components, and group ability patterns rather than on detection of individual aberrant response patterns (Burke, 2009).

Administrative quality monitoring methods refer to analytical methods that can be used to detect any type of degradation in the quality of standardized conditions for a testing program. Since there are many computer-based examination platforms in use today, changes can inadvertently be introduced when a new version of software is used or when new hardware devices are installed. For example, if a new test driver version inadvertently allows examinees to access an online calculator that was not permitted on previous versions of the software, the feature could create an unfair advantage for examinees who take the newer software version. Forensic analysis methods that monitor trends in average item latencies (average examinee time spent on an item) would likely identify short latencies and could be tracked to a change extraneous to exam content.

The Concept of Reasonableness

Reasonableness is the basis for most data forensics analyses. *Reasonableness* is a term used by psychometricians and data specialists to refer to the tendency of regular testing events to produce results that are consistent with historical data. Reasonableness is usually used as a final quality control check by measurement experts to ensure that results from chains of analyses or long data control steps are, at a minimum, homogeneous with previous results. For example, when testing programs do complex equating procedures, a final step often requires a senior measurement expert to compare the proposed equating solution for a test administration to previous solutions and to decide whether the solution is close enough to be considered "reasonable" for the testing program. An unreasonable result often is a red flag to the data analysts and requires a thorough review or even replication of the stream of equating results. In a similar way, data forensics capitalizes on portions of data that are not reasonable or data that show some characteristic of outliers compared to historical data.

Importance of Baseline Data

To determine whether current data are reasonable, it is usually important to have some baseline data for comparison. At times the baseline data can be extrapolated from a very small sample or from an a priori notion of expected baseline data. For instance, if data from a particular test are not available, then information from similar testing programs may be used to form a prior understanding of the baseline data for the testing program under investigation. In fact, formal Bayesian prior distributions could be used to augment the baseline data until sufficient data become available. Baselines can be established for test characteristics, item characteristics, examinee groups, or individual examinees. For instance, a number of testing programs use retake analysis indicators to quickly identify evidence that indicate very large differences in scores between retakes. Examinees whose results show large score differences between proximate retakes on the same examination might indicate cheating behavior, identity falsification, or systematic problems with the exam (Impara, Kingsbury, Maynes & Fitzgerald, 2005). Such results are sometimes automatically placed on hold and subject to careful psychometric review until a satisfactory explanation for such unlikely discrepancies has been made.

Formulating the Right Question

Many issues surrounding data forensics are related to security and quality concerns for a testing program. It is important first to formulate the types of questions that may be answered by the data. Many of these types of questions cannot be answered with complete confidence, but only with varying degrees of likelihood. For instance, a common concern is item or test form compromise. Are any of the items in the item bank or on a testing form known by large numbers of examinees prior to taking their exams? Direct evidence of preknowledge may be impossible to prove; however, likelihood statements may be made about an item pool based on available samples. For instance, in a large-scale

program, it is possible that hundreds of examinees may have unfairly gained preknowledge of items, yet the testing population could be so large as to make this group impossible to detect with typical global form and item-level analyses. Instead, regional or other demographic control variables might be used along with blocking or analysis of covariance methods to achieve enough sensitivity in the analyses for detecting differences in the population (Impara et al., 2005).

Another common issue might relate to verification of examinee identity (as in the retake example discussed above). Collusion is another major concern of test sponsors, especially with the proliferation and miniaturization of many electronic communications devices today. Direct answer copying may still pose a risk, but with the widespread use of computerized testing, collusion may refer more generally now to item harvesting attempts or test session covert communications to gain a material advantage on a particular exam (Gross, 2003).

Defensibility Issues

One difficulty in identifying cheating behavior lies in the fact that some patterns, although highly improbable, are rarely impossible to rule out. Unless a testing program has explicit policies and guidelines laid out in advance, it is difficult and sometimes unwise to invalidate scores simply on the basis of improbable response patterns. These issues must be considered carefully in coordination with measurement experts, policy makers, and legal counsel for an organization. Some options include enacting policies that release scores only if those scores show appropriate fit to the measurement model. The reasons for misfit may not need to be stated, and organizations can avoid pointing fingers at individuals or their motivations and can simply rely on the notion that severe misfit to the model does not produce valid scores.

Legal defensibility often depends on having an enforceable and valid test use agreement in place between examinee and testing program before the start of an examination (Foster, Maynes & Hunt, 2008). Examinees can be expected to consent to such requirements as not removing or disclosing confidential material (e.g., items) that they see on an exam, and not selling such material. Test sponsors may also stipulate that test scores may be revoked if forensic analysis reveals a high likelihood that the score is not valid. Such cases have held up legally as long as an enforceable legal agreement is in place (e.g., Cizek, 2004; Foster et al., 2008).

Summary

Most full service testing organizations today offer forensic data analysis products. Some companies even specialize in data forensics, security audits, and quality control methods. Increasingly, test sponsors are using these products to maintain a protective barrier around their tests and to guard against test compromises, incursions into their testing systems, or degradations in examination quality. Surveys in future trends in the testing industry suggest that forensic methods will become more widely used and improved security and identity verification will be emphasized in the near future as test sponsors focus on security and quality assurance issues for their testing programs (Becker & Pascal, 2010). The field of data forensics is destined to grow and to become more sophisticated as industry needs broaden. Forensics is a growth area for the future of the industry and an important area for future psychometric research.

References

Becker, K. & Pascal, P. (2010, February). Predicting the future of testing: A Delphi study. Paper presented at the Annual Meeting of the Association of Test Publishers. Orlando, FL.

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

Cizek, G.J. (2004, April). Protecting the integrity of computer-adaptive tests: Results of a legal challenge. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Foster, D., Maynes, D. & Hunt, B. (2008). Using data forensic methods to detect cheating. In C. L. Wild & R. Ramaswamy (Eds.), *Improving Testing* (pp. 305–321). New York: Lawrence Erlbaum.

Gross, L. (2003). Psychometric matters. *CLEAR Exam Review*, 14, 15–16.

Impara, J.C., Kingsbury, G., Maynes, D. & Fitzgerald, C. (2005, April). Detecting cheating in computer adaptive tests using data forensics. Paper presented at the annual meeting of the National Council of Measurement in Education, Montreal, Canada.