Technology-Enhanced Assessment of Talent

Nancy T. Tippins
Seymour Adler
Editors

Foreword by
Allen I. Kraut
The Professional Practice Series

The Professional Practice Series is sponsored by The Society for Industrial and Organizational Psychology, Inc. (SIOP). The series was launched in 1988 to provide industrial and organizational psychologists, organizational scientists and practitioners, human resources professionals, managers, executives and those interested in organizational behavior and performance with volumes that are insightful, current, informative and relevant to organizational practice. The volumes in the Professional Practice Series are guided by five tenets designed to enhance future organizational practice:

1. Focus on practice, but grounded in science.
2. Translate organizational science into practice by generating guidelines, principles and lessons learned that can shape and guide practice.
3. Showcase the application of industrial and organizational psychology to solve problems.
5. Stimulate research needed to guide future organizational practice.

The volumes seek to inform those interested in practice with guidance, insights and advice on how to apply the concepts, findings, methods, and tools derived from industrial and organizational psychology to solve human-related organizational problems.
Contents

Foreword by Allen I. Kraut, Baruch College/Kraut Associates xiii
The Editors xvii
The Contributors xix
Preface xxxv
Acknowledgments xxxvii

1. Overview of Technology-Enhanced Assessments 1
Nancy T. Tippins

Section One: Measurement and Implementation Issues
in Technology-Enhanced Assessments 19

2. Foundations for Measurement 21
John C. Scott and Alan D. Mead

3. Implementing Assessment Technologies 66
Douglas H. Reynolds

4. Cheating and Response Distortion on Remotely
Delivered Assessments 99
Winfred Arthur, Jr., and Ryan M. Glaze

5. Computerized Adaptive Testing 153
Rodney A. McCloy and Robert E. Gibby

6. Applicant Reactions to Technology-Based Selection:
What We Know So Far 190
Talya N. Bauer, Donald M. Truxillo, Kyle Mack, and Ana B. Costa

7. International Issues, Standards, and Guidelines 224
Dave Bartram

Section Two: Case Studies of Technology-Enhanced
Assessments 251

8. Web-Based Management Simulations: Technology-
Enhanced Assessment for Executive-Level Selection
and Development 253
Terri McNelly, Brian J. Ruggeberg, and Carrol Ray Hall, Jr.
Contents

9. Bridging the Digital Divide Across a Global Business:  
Development of a Technology-Enabled Selection System for Low-Literacy Applicants  
Adam Malamut, David L. Van Rooy, and Victoria A. Davis  
267

10. Promotional Assessment at the FBI: How the Search for a High-Tech Solution Led to a High-Fidelity Low-Tech Simulation  
Amy D. Grubb  
293

11. Innovation in Senior-Level Assessment and Development: Grab ‘Em When and Where You Can  
Sandra B. Hartog  
307

12. Case Study of Technology-Enhanced Assessment Centers  
Rick Hense and Jay Janovics  
324

13. Video-Based Testing at U.S. Customs and Border Protection  
Jeffrey M. Cucina, Henry H. Busciglio, Patricia Harris Thomas, Norma F. Callen, DeLisa D. Walker, and Rebecca J. Goldenberg Schoepfer  
338

14. Going Online with Assessment: Putting the Science of Assessment to the Test of Client Need and 21st Century Technologies  
Eugene Burke, John Mahoney-Phillips, Wendy Bowler, and Kate Downey  
355

15. Implementing Computer Adaptive Tests: Successes and Lessons Learned  
Mike Fetzer and Tracy Kantrowitz  
380

Michael J. Zickar and Christopher J. Lake  
394

17. Concluding Comments: Open Questions  
Seymour Adler  
418

Indexes  
Name Index  
Subject Index  
437  
445

Foreword

Who would have thought that one of the most important “laws” affecting personnel assessment would be Moore’s Law? About forty-five years ago, Gordon Moore, a co-founder of Intel, described his expectation that the cost of computer chips would be cut in half about every two years for the foreseeable future. This proposition, often described as Moore’s Law, turned out to be largely true, cutting the cost of computing and related processes by half every two years since then. These drops in cost have made computing power cheap and easily available in many devices such as personal computers, laptops, and cell phones.

Together with other advances in technology, such as fiber optics and space satellites, a vast array of devices are now available for personnel assessment and selection purposes that could not have been imagined just two decades earlier. Younger practitioners in human resources and industrial/organizational psychology must be forgiven if they think that such devices and applications were always with us. In fact, the computing power of a BlackBerry cell phone today exceeds the power of the computer used to guide the first spaceship to land on the moon.

Technology and what it allows us to do in the field of assessment have unfolded incredibly fast. The classic Personnel Testing by Robert M. Guion (1965) does not even contain the word “computer” in its index. Catching up, Guion and Highhouse’s 2006 book on the same subject does have “computerized testing” in the subject index, showing it was mentioned on four pages.

One of the earliest recognitions of the potential use of computers and allied devices in psychological assessment appeared in 1998 in a book chapter by Rob Silzer and Richard Jeanneret, two highly experienced and creative practitioners. They foresaw “the broader use of computers and information technology in the assessment process . . . it will also be extended to simulations and exercises and to multi-rater questionnaires.” Hesitantly, they noted “one
Chapter Four

CHEATING AND RESPONSE DISTORTION ON REMOTELY DELIVERED ASSESSMENTS

Winfred Arthur, Jr.,* and Ryan M. Glaze

In this chapter we provide the reader with a succinct review of the state of the literature on cheating and response distortion on remotely delivered ability and knowledge, and nonability and noncognitive assessments, respectively. This review and associated discussion is organized around five questions: (1) What is cheating and response distortion? (2) How pervasive and extensive are they? (3) How are they detected and how effective are said detection techniques? (4) What does one do with either the information, cheaters and dissimulators, or both, once they have been detected? and (5) How does one deter these behaviors (cheating and response distortion) and how effective are

---

*Correspondence concerning this chapter should be addressed to Winfred Arthur, Jr., Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843-4235. Email should be sent to w-artur@neo.tamu.edu.
said deterrence techniques and methods? (It is important to note that we do not discuss the effect of cheating and response distortion on the psychometric properties of remotely delivered assessments; for a review of these issues, the interested reader is referred to Chapter 2 by Scott and Mead on the Foundations for Measurement in this volume.) We conclude the chapter with a few summary statements and a brief discussion of directions for future research. One such important summary statement is that the proctored versus unproctored delivery of assessments has more profound effects and implications for ability and knowledge tests than it does for noncognitive and nonability measures primarily because proctoring is not a means for controlling for response distortion on the latter type of assessments. Another equally important summary statement is that the amount of research on unproctored Internet-based ability and knowledge tests is very limited, and given the concerns about cheating on unproctored Internet-based ability and knowledge tests, it is definitely not commensurate with the use of these types of assessments in practice.

For the purposes of this chapter, we use the terms “remotely delivered assessments” and “unproctored Internet-based tests” synonymously and interchangeably. Within this nomenclature, collectively, remotely delivered assessments are typically characterized by the absence of a human proctor to ensure that test-takers follow and abide by (ethical) testing rules and standards. This characteristic justifiably raises misgivings about the accuracy and validity of test scores obtained from this mode of testing because in the absence of a proctor or proctoring, there are concerns that test-takers, left to their own devices, are likely to engage in illicit activities to present themselves in as favorable a light as possible—especially in the context of high-stakes testing. Thus, the opportunity for malevolent behaviors is a major psychologically-based source of construct irrelevant variance in that these behaviors (may) produce test scores that do not accurately reflect the test-taker’s standing on the constructs of interest.

We define “malevolent behavior” as deliberately falsifying or misrepresenting one’s responses on a test, assessment tool, or device in an attempt to distort one’s true standing on the constructs of interest. Malevolent behaviors may take one of two forms—cheating or response distortion. Cheating is associated with ability and knowledge tests and entails the use of illicit aids to obtain and produce the keyed or correct answers to test items. Cizek (1999) defines cheating as “an attempt, by deceptive or fraudulent means, to represent oneself as possessing knowledge [or ability]. In testing specifically, cheating is violating the rules” (p. 3). Response distortion, on the other hand, is associated with noncognitive measures and refers to deliberately falsifying one’s responses to self-report items, taking the form of faking, impression management, and other forms of non-truthful responding with the goal of presenting oneself in as (socially) favorable a light as possible (Paulhus, 2002). Consequently, in the present chapter, we use the term “cheating” to refer to malevolent behaviors on ability and knowledge tests and “response distortion” to refer to malevolent behaviors on noncognitive and nonability measures.

**Cognitive Ability and Knowledge Tests**

Ability and knowledge tests and measures are characterized, and thus are distinguishable from noncognitive self-report measures, by the fact that they have demonstrable incorrect and correct or best answers. To this end, in situations in which test security is a concern, proctoring is the primary means by which cheating is curtailed. Cheating on knowledge and ability tests may take the form of employing illicit aids such as cheat sheets, calculators and dictionaries, surrogate test-takers (for example, a smart friend), or preknowledge of test items. Cheating is commonly inferred by either direct observation or statistically. In the context of unproctored Internet-based tests, the latter approach is more germane and typically calls for a repeated administration of the test—the first under unproctored conditions and the second under proctored conditions. Differences in performance on the two administrations are then used to make inferences about the absence or presence of cheating. Needless to say, said repeated administrations may not always be feasible or practical.
Amount of Cheating

Although the number of published studies on the prevalence of cheating in employment testing and settings is very limited, cheating in educational settings is very well documented and seems to be quite widespread as well. For example, reviews by Cizek (1999) and Whitley (1998) summarized research indicating that 50 percent or more of all college students reported cheating on an exam at least once during their college education. Indeed Chapters 2 and 3 of Cizek’s book present a trove of information on the frequency and perceptions of cheating and common methods of cheating; and Chapter 4 discusses cheating in postgraduate and professional contexts. Although the focus of most of the writing and data has been on cheating in academic contexts (primary, secondary, higher, and postgraduate and professional education), suffice it to say that, even with the limited data that are available it is not unreasonable to assume that if unchecked the proclivity to cheat on employment tests and exams is probably just as high as that which has been observed in educational settings.

For instance, it has been reported that 45 percent of U.S. job applicants falsify their work histories (Automatic Data Processing, Inc., 2008). Consistent with this, both Hense, Golden, and Burnett (2009) and Beaty, Fallon, Shepherd, and Barrett (2002) reported higher scores on the unproctored versions of their tests. The respective standardized mean differences for these two studies were 0.32 and 0.52. Furthermore, using a 3 standard error of measurement (SEM) operationalization of cheating, Beaty, Fallon, Shepherd, and Barrett identified 9 percent of the top seventy-five scorers as having cheated on the unproctored administration. Likewise, using a design in which applicants completed an unproctored Internet-based cognitive ability test first as job applicants and then as research participants (average retest interval of 429 days), Arthur, Glaze, Villado, and Taylor’s (2010) SEM operationalization of cheating identified 7.77 percent of their sample as having cheated with these individuals being distributed across the entire range of test scores. It is also worth noting that from one perspective, their 7.77 percent may represent an underestimate of the prevalence of cheating since the cognitive ability test was designed to be very speeded to prevent cheating. In addition, since the (research-based) retest was on a volunteer basis and consequently, via self-selection, it may have had a larger proportion of individuals who did not cheat on the first, high-stakes assessment. Finally, in a finding in education that has resonance for unproctored Internet-based testing, Whitley’s (1998) review found that an honor code coupled with subsequent unproctored exams resulted in higher levels of self-reported cheating on the unproctored exams.

To summarize, ability and knowledge tests are susceptible to cheating for several reasons. First, in organizational, educational, and other high-stakes settings (for example, professional certification or licensure), test-takers’ scores play an important role in whether they will be hired, admitted, or otherwise obtain or achieve their desired outcome. Second, the transparency and valence of ability and knowledge test items are clear—that is, these test items have demonstrable incorrect and correct or best answers. So, unlike personality measures, the desired response is a matter of fact. Consequently, although additional research is needed to further document its prevalence and extent, it is reasonable to conclude that cheating can and does occur in applicant and other high-stakes organizational testing and that it is likely to be exacerbated with the use of unproctored remotely delivered assessments since the absence of proctors creates a permissive environment for cheating. We obviously do not claim or intend to imply that proctoring testing is immune from cheating—to the contrary, data such as that reviewed and summarized by Cizek (1999) and Whitley (1998) would suggest otherwise. Since proctoring is the primary means by which rule compliance is enforced, it is not unreasonable to posit that in the absence of proctors those who are motivated to do so may use surrogates and advisors and illicit reference materials, calculators, and dictionaries in an effort to increase their ability or knowledge test scores. The resultant expectation is that the presence of cheating will result in elevated test scores.
Differences in Cheating in Proctored and Unproctored Internet-Based Settings

Two noteworthy issues in the discussion of differences in cheating in proctored and unproctored Internet-based settings are the methods and frequency of cheating. Methods of cheating that would appear to be common to both proctored and unproctored Internet-based tests include cheating by taking, giving, or receiving information from others, and cheating through the use of forbidden materials and information. Cheating via the use of surrogate test-takers would also seem to be common to both although one could reasonably speculate that it is easier to accomplish this with unproctored Internet-based tests than proctored tests. Finally, Cizek (1999) describes a category of cheating methods that involve “taking unfair advantage of the person(s) giving the test or the circumstances of the testing process. For example, students can take advantage of vague, ambiguous, or uncontrolled test administration protocols, their instructor’s willingness to help...” (p. 48). As described, this category of cheating methods would seem to be particularly more germane to proctored than unproctored settings.

A methodological aspect of cheating that characterizes only unproctored Internet-based testing pertains to attacks on the technology used to support Internet-based testing. Thus, pirate and hacker attempts to access test content, scoring keys, and test score data and, subsequently, making them available to test-takers (Burke, 2009), is a threat that is less germane in proctored non-Internet-based settings. Another aspect of unproctored Internet-based testing that may differentiate it from proctored testing is the international scope of the former, which derives primarily from its often touted advantage of being a “test anywhere-test anytime” method. With the increasing emphasis on corporate globalization, and the concomitant globalization of organizations’ human resource management systems and practices, the use of Internet-based assessments is consonant with this business model. However, an unintended but not surprising result of this is that the security threats and concerns that confront a specified Internet-based assessment tool are consequently also international in scope. So, for instance, as part of an effort to maintain the integrity of their Internet tests, SHL conducts web patrols to detect test security breaches and test fraud; and reflective of the international scope of security threats, Burke (2009) reports that of the eighteen (out of thirty) websites that SHL detected and classified as high-risk in an eighteen-month period, four were sites that were operating in England and fourteen were in China.

Concerning frequency and prevalence, because proctored settings, by their very nature, use human observation and supervision to curtail, prevent, and detect these cheating methods, one would expect the prevalence of cheating—at least overt cheating—to be lower in proctored settings. Thus, in proctored settings, one would expect cheating to be associated with the extent to which test-takers can or believe that they can outwit human proctors, who because they are neither perfect or infallible, are unable to prevent or detect all cheating attempts. Thus, said cheating attempts would also be expected to be more covert and clandestine. In contrast, one would expect cheating in unproctored settings to be more open and overt and less clandestine. Therefore, the logical inference is that cheating in unproctored settings may be more prevalent and frequent, which is consonant with Whitley’s (1998) finding that students were more likely to cheat when they thought there was relatively little risk of being caught and when they anticipated large rewards for success on the test.

In summary, proctored and unproctored settings share some common cheating threats. Similarly, using several telling examples, Drasgow, Nye, Guo, and Tay (2009) make a compelling case that in reference to cheating, proctored assessments cannot necessarily be considered the “gold standard”. Nevertheless, because the use of human observers and proctors is the primary means by which several threats and specified cheating methods are prevented and detected, one would expect the use of cheating methods in unproctored settings to be more overt and less clandestine, and concomitantly, one would also expect the frequency and prevalence of cheating in unproctored settings to also be higher. There are also aspects of cheating, such as attacks on the test hosting server or site and the global and international nature of cheating-related issues and concerns, that are more applicable to unproctored Internet-based testing. However, the critical question is that, although cheating does occur, is there
any clear, overwhelming empirical evidence that conclusively demonstrates that the amount of cheating is more than trivial? And furthermore, is it qualitatively more than that which occurs with proctored tests? As Drasgow, Nye, Guo, and Tay (2009) emphasize, "Cheating on proctored tests is not difficult, and there is evidence to suggest that it may be pervasive in some situations" (p. 48). Based on our review of the literature, at the present time it seems we simply assume—and maybe justifiably so—that cheating on unproctored tests is more prevalent. However, we must acknowledge that it is just that, a reasonable supposition, and that there is very limited empirical research that has investigated and conclusively and overwhelming demonstrated this.

Detection

Although there is agreement about the need to be concerned about cheating on unproctored Internet-based tests, given its elusive nature, it is difficult to directly measure this behavior. Thus, in both research and applied settings, rarely do we actually see or observe applicants or test-takers cheating, nor is there typically direct evidence that they did. As a result, the techniques used to detect cheating are not directly behavioral, but instead, cheating is determined and operationalized in terms of inferences that are made from test scores. In the absence of directly observing test-takers cheating, techniques for detecting cheating in unproctored Internet-based testing take the form of (1) statistical detection, (2) score comparison and verification testing, and (3) technological detection.

Statistical Detection

Statistical detection methods assess (1) the similarity of (pairs of) scores or response patterns (of errors and correct responses) or (2) deviations from some known or expected distribution of scores which may be probabilistic or actual (retest) scores. Because they originate from educational testing, statistical detection methods focus primarily on answer copying as a particular source of cheating. In this regard, Cizek (1999; see also Hanson, Harris, & Brennan, 1987; Sauer, 1960) notes the distinction between chance and empirical statistical detection methods; adding that, over time, detection methods have moved away from empirical to chance methods. Chance methods "compare an observed pattern of responses by a pair of examinees (one or both of whom are suspected to be cheating) to a known distribution, such as the binomial or standard normal distribution. Empirical methods [on the other hand] compare the probability of an observed pattern of responses by a pair of examinees to a distribution of values derived from other independent pairs of students who took the same test. Distributions of empirical indices for suspected copiers are compared to distributions of statistics obtained under conditions where cheating could not have occurred" (Cizek, 1999, p. 137). Regardless of the specific approach used, individuals with a large index (for example, B, K, g) are subsequently flagged as having cheated. The interested reader is referred to Cizek (1999) for a more detailed presentation of these statistical detection methods that apply and are used almost exclusively for detecting whether a test-taker copied from another. In summary, statistical detection methods are mostly used in educational testing settings and may have limited applicability to unproctored Internet-based testing as used in organizational and employment testing contexts—with the military being an exception to this. A final reason for its limited applicability in (civilian) organizational settings is that statistical detection is a post hoc approach that requires relatively large sample sizes that are not present in most organizational settings.

Score Comparison and Verification Testing

In unproctored Internet-based testing of the sort that is commonly used in organizational settings, copying the answers of an adjacent test-taker is less of a concern; the focus is more on rule violations pertaining to the use of illicit aids such as calculators and dictionaries, preknowledge of test items, and the use of surrogate test-takers. So, in an effort to utilize unproctored Internet-based testing and maintain test utility and validity, the use of proctored retesting to verify and confirm unproctored test scores has been advocated (International Test Commission [ITC], 2005; Tippins, Beaty, Drasgow, Wade, Pearlman, Segall, & Shepherd, 2006). Proctored retesting can take the form of a full length retest (repeating the original test or an alternate form of
equal length) or an abridged retest which uses a shorter form of the original test. Regardless of the specific approach used, score differences between the two administrations—that is, when the proctored score is lower than the unproctored score—are then used to make inferences about cheating.

Psychometric theory and research indicates that retesting is generally associated with increases in test scores. For instance, Hausknecht, Halpert, Di Paolo, and Moriarty-Gerrard's (2007) meta-analytic results indicate that test-takers increase their retest scores both with coaching \((d = 0.64)\), and without coaching \((d = 0.21)\). Consequently, within some standard error of measurement, to the extent that a test-taker's proctored retest score is lower than his or her first unproctored score, the unproctored score is considered to be suspect because, psychometrically, it is expected to be the same or higher, not lower, than the first test score. In summary, the primary issues in verification testing are (1) score equivalence and (2) the techniques and thresholds that one uses to determine that the unproctored and proctored scores are different and, thus, warrant the suspicion or conclusion of cheating.

A number of empirical studies have investigated these very issues. Hense, Golden, and Burnett (2009) report the results of a between-subjects design in which applicants who took a video-based job simulation when it was first rolled out as a proctored test obtained lower scores than those who took it when it was later reintroduced as an unproctored Internet-based test \((d = 0.52)\). As the authors acknowledge, although the higher scores on the unproctored version are consistent with a cheating hypothesis, in the absence of random assignment, alternative explanations such as different testing conditions, locations, and test-takers for the two versions cannot be ruled out. In addition, group mean differences do not identify specific individuals who may have cheated; they merely indicate that, at the group level, unproctored test scores are higher than proctored test scores.

In a study that is less susceptible to the preceding methodological threats, Beaty, Fallon, Shepherd, and Barrett (2002) report the results of a multi-stage assessment process in which applicants who were not screened out on the basis of minimal requirements ineligibility took a speeded (twelve minutes for fifty-four items) unproctored Internet-based cognitive ability test. The top seventy-five scorers from the unproctored administration were then invited to complete a parallel form of the test under proctored conditions. The average retest interval was four weeks. The results of this within-subjects design were consistent with the cheating hypothesis—the unproctored scores were higher than the proctored scores \((d = 0.51)\). In addition, using a 3 SEM operationalization, six individuals (9 percent) had proctored scores that were 3 SEMs lower than their unproctored scores and so their unproctored scores were flagged as suspect.

Arthur, Glaze, Villado, and Taylor (2009, 2010) also used a within-subjects design. However, they compared two unproctored conditions such that on the first administration, test-takers completed the unproctored Internet-based speeded cognitive ability test (twenty minutes for 120 items) as job applicants (high-stakes) and on the second administration volunteered to retake the same unproctored test for research purposes (low-stakes). The average retest interval was 429 days. However, unlike Hense, Golden, and Burnett (2009) and Beaty, Fallon, Shepherd, and Barrett (2002), Arthur, Glaze, Villado, and Taylor's findings were more consonant with a psychometric than a cheating explanation since the retest scores (although low-stakes) were higher than the initial high-stakes scores \((d = 0.36)\). In addition, using a 1 SEM operationalization, only 7.77 percent of the sample (twenty-three out of 296) had Time 1 scores that were 1 SEM lower than their Time 2 scores. Arthur, Glaze, Villado, and Taylor (2010) attributed this low incidence of cheating to the very speeded nature of the test, which was intentionally designed as such to increase the temporal costs associated with and, hence, deter cheating. However, the authors also acknowledge that, although speededness may deter some forms of cheating, it does not curtail the use of surrogate test-takers. Finally, since the (research-based) retest was on a volunteer basis and, thus, via self-selection, it may have had a larger proportion of individuals who did not cheat on the first high-stakes assessment.

In a two-step testing process in which candidates first completed an unproctored Internet-based speeded perceptual speed and accuracy test and then retested on a parallel form under
proctored conditions on average a month later (both high-
stakes), Nye, Do, Drasgow, and Fine (2008) obtained results that
were quite similar to Arthur, Glaze, Villado, and Taylor (2010).
Specifically, the proctored retest scores were higher than the
unproctored initial test scores ($d = 0.22$). In addition, using a
regression-based approach, Nye, Do, Drasgow, and Fine iden-
tified only four individuals (out of 856) who showed differences
larger than the 1.96 cutoff.

The research reviewed above used different approaches—
namely SEM- and regression-based—to infer whether differences
between unproctored and proctored test scores were indicative
of cheating or not. Along these lines, Guo and Drasgow (2009)
present a $Z$-test and a likelihood ratio test as alternative means of
comparing the consistency of performance across unproctored
and proctored testing conditions and subsequently, identifying
the dishonest test-taker. Based on the results of a simulation
study, Guo and Drasgow conclude that (1) both test statistics
have a high power to detect cheating at low Type I error rates
and (2) compared to the likelihood ratio test, the $Z$-test is more
efficient and effective.

The use of score comparisons and verification testing raises a
number of additional noteworthy issues. First, proctored retesting
obviously adds additional steps to the recruitment and testing
process and thus, to some extent diminishes the efficiencies and
cost effectiveness advantages often touted in support of unpro-
tored Internet-based tests (that is, “test anywhere-test anytime”).
Second, as Pearlman (2009) notes, retesting paradoxically con-
stitutes “a tacit admission by the organization that UIT results
cannot be relied upon, which may have negative indirect con-
sequences” (p. 15).

Third, as previously alluded to, retesting raises interesting
issues about determining whether the unproctored and pro-
tored test scores are from the same person. From an applied
perspective, one could argue that the importance of this depends
on which test score is intended to be the score of record. Thus,
the criticality of determining whether the two scores are from the
same person or not is largely a function of whether the unpro-
tored score is the score of record, such that it is less of an issue
and concern if the proctored score is the score of record.

Fourth, related to the preceding is the question of which of
the two test scores should be used as the score of record from
a decision-making perspective—the first (unproctored) or the
second (proctored) score, or some combination of the two?
Although this would appear to be a critical issue with important
psychometric and applied implications (for instance, Arthur,
Glaze, Villado, & Taylor [2010], Beaty, Fallon, Shepherd, &
Barrett [2002], and Nye, Do, Drasgow, Fine [2008] reported
retest correlations of .78, .41, and .63, respectively, between
the initial and retest test scores of their tests), we were unable to
locate any studies that addressed or investigated the issue of
which score to use as the operational score. Nevertheless, Burke
(2009) notes that SHL uses the first score as the score of record
and treats the verification test score (“short tests administered
in a proctored setting” [p. 36]) in a manner that is analogous
to a fake-good check on a personality measure. However, to the
extent that the second (proctored) administration uses either
the same or an equivalent test, a reasonably good case could be
made for using that as the score of record and the first as just an
initial screen—but in the absence of any empirical research, it is
difficult to decide which approach is better and subsequently,
to make any firm recommendations.

A fifth issue is that of measurement equivalence. For instance,
Lievens, Reeve, and Heggestad (2007) reported that proctored
retesting is associated with changes in the factor structure of ability
tests such that test scores based on a retest of general mental
ability are less saturated with general mental ability and, subse-
quently, less predictive of grade point average than initial test
scores. These results suggest that the use of proctored retesting
may threaten the construct-related and criterion-related validity
of test scores gathered using this approach. Consequently, this is
an issue that should be considered when deciding whether to use
the unproctored (first administration) or the proctored (second
administration) scores as the operational assessment scores.

In summary, the empirical research reviewed above indi-
cates that verification testing and score comparisons are viable
approaches for detecting score differences that may subsequently
be interpreted as cheating. In operational, as opposed to research
terms, this will entail retesting that is also high-stakes using either
repeatedly before the start of the test to establish a keystroke pattern. The system then periodically requests the same password to be retyped, and the typing pattern is compared to the established keystroke pattern. If the established keystroke pattern cannot be reproduced, the inference is that another individual is now taking the test. Real-time data forensics is a technique of flagging test-takers who display suspect item response or response latency patterns. For example, if a test-taker correctly answers several difficult questions then incorrectly answers several easy questions, the program flags the test for future review. Keystroke monitoring is a technology that allows the system to record test-takers’ attempts to use illicit keys (for example, alt-tab, print screen) and alert a remote proctor or suspend or stop the test. Other security options include software programs that limit test-taker’s access to available information (browser lockdown, operating system and Internet access control). Web patrols may also be used to detect Internet sites that compromise test materials (Burke, 2009).

In summary, although technological detection techniques hold some promise, they have some disadvantages, including cost, invasiveness, and applicant reactions to the testing process. And some even require some level of human proctoring, albeit technologically mediated. In addition, some of these methodologies are currently state-of-the-art and may consequently not be a viable option (from a cost and expertise perspective) in most small-scale testing programs. Hence, it is not too surprising that they are currently not as widely used and as commonplace as, for example, verification retesting. Nevertheless, it is important to note that technology changes rapidly and it is likely that new, and perhaps more available and cheaper, means of detecting (and preventing) cheating will be developed. Finally, it should also be noted that there is a dearth of research that empirically investigates the effectiveness of these techniques.

Once Detected, What Does One Do with and About Cheats and Perpetrators?

The endeavor of trying to detect cheating begs the question, "What does one do with and about cheats and perpetrators once they have been detected?" This ostensibly simple question raises
several interrelated issues. The first pertains to “What constitutes positive evidence of cheating?” Clearly, simple score differences, even after taking into account some specified amount of measurement error (which encapsulates sources of error variance due to factors such as illness and practice), is unlikely to meet the evidentiary threshold; a condition that is equally applicable to most of the technological detection methods as well. Indeed, Cizek (1999) discuss several instances when academic administrators have been very hesitant to press academic dishonesty charges when human proctors have accused students of cheating on exams on the basis of their having observed them doing so—it would seem that even “charges” based on only observation may not be sufficient. Ironically, even in large-scale academic testing such as those undertaken by ACT and ETS, statistical evidence is brought to bear only when some other trigger (for example, observation) provides a strong reason for flagging cases for subsequent statistical analysis (Cizek, 1999). Consonant with this, in describing the ITC’s (2005) position on this issue, Bartram (2009) notes that “[i]n a proctored environment, there needs to be positive evidence of cheating rather than just circumstantial evidence” (p. 13).

A second related issue is that of false positives or identifications. Within this context, it is important to emphasize that none of the prevailing detection methods actually detect cheating. They only provide evidence or data that permit inferences about the presence of cheating. Thus, the conclusion that cheating has occurred on the basis of these methods is almost always probabilistic and requires an inference. Given these limitations, confronting suspected cheaters and perpetrators, although necessary in academic testing, is neither required nor necessary in employment testing. However, an employment-related decision has to be made on the basis of the specified test scores. In verification testing, this decision is probably moot if the proctored retest is used as the operational score or record. On the other hand, it is a much bigger concern if the unproctored initial test score is the operational score.

So, if we do not have to confront them, can we fail them for having cheated? Whereas the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) does not currently speak specifically to unproctored Internet-based testing, it has several standards that serve as professional guidelines concerning “test irregularities” and how to deal with them. Standards 8.11, 8.12, and 8.13 appear to be particularly germane. For instance, Standard 18.11 states that:

“In educational testing programs and in licensing and certification applications, when it is deemed necessary to cancel or withhold a test-taker’s score because of possible testing irregularities, including suspected misconduct, the type of evidence and procedures to be used to investigate the irregularity should be explained to all test-takers whose scores are directly affected by the decision. Test-takers should be given a timely opportunity to provide evidence that the score should not be canceled or withheld. Evidence considered in deciding upon the final action should be made available to the test-taker on request.” (p. 89)

In contrast to the current Standards, the ITC Guidelines (ITC, 2005) speak specifically to unproctored Internet-based tests and the pertinent guidelines that address the issue of cheating are Guidelines 49, 44, and 45. For instance, Guideline 45.3 explicitly stipulates the need for verification testing noting that:

“For moderate and high stakes assessment (for example, job recruitment and selection), where individuals are permitted to take a test in controlled mode (i.e., at their convenience in non-secure locations), those obtaining qualifying scores should be required to take a supervised test to confirm their scores. Procedures should be used to check whether the test-taker’s original responses are consistent with the responses from the confirmation test. Test-takers should be informed in advance of these procedures and asked to confirm that they will complete the tests according to instructions given (for example, not seek assistance, not collude with others, etc). This agreement may be represented in the form of an explicit honesty policy which the test-taker is required to accept.” (pp. 20–21)

So, like the Standards, the ITC Guidelines also stipulates that test-takers should be informed in advance of the retesting procedures.
In conclusion, in our opinion, it would seem that the most efficacious and straightforward way to deal with cheats and perpetrators is to have a verification retest and use this score as the score of record. Test-takers should also be informed in advance of this along with the fact that the unproctored test is just an initial screen. This approach preempts the need to confront test-takers about having cheated.

**Deterrence**

Given the practical, ethical, and legal challenges associated with not only the detection of cheating, but also what do with and about applicants suspected of cheating, the proverb, “an ounce of prevention is better than a pound of cure” is particularly germane here. Approaches to deter cheating can be conceptualized as having one of two foci—(1) to discourage cheating attempts and (2) to make it more difficult to engage in cheating. Approaches that fall into the first category include monitoring and the saliency of other detection techniques (including web patrols), and warnings and threats. The second category consists primarily of test design characteristics and features such as the use of multiple test forms, computerized adaptive tests, and speeded tests.

**Monitoring**

Because we define an unproctored Internet-based test or remotely delivered assessment as one that does not entail the direct monitoring or supervision of a human proctor in the same physical setting or location as the test-taker, our focus here is on electronic monitoring primarily in the form of webcams, keystroke analytics, and keystroke monitoring. Foster (2009) describes new technology-enhanced ways of monitoring remotely delivered assessments that do not require the physical presence of a proctor in the same room with the test-taker. But by virtue of needing a human to monitor the data obtained from these systems (for example, monitoring webcam monitors or other data), these technology-based monitoring systems are ultimately susceptible to the human-related concerns and limitations that characterize traditional proctored tests.

It is also worth noting that, although they may not be designed or intended to specifically do so, it would seem that the saliency of detection techniques such as biometric identification, web patrols, and webcams, coupled with test-takers’ beliefs about their effectiveness, are also likely to serve a deterrent function as well. That being said, it should be noted that there is a paucity of empirical research investigating the effectiveness or efficacy of these techniques in deterring or discouraging cheating.

**Warnings and Threats**

Like the saliency of detection techniques, warnings and threats may also serve to discourage and deter cheating. Conceptually, warnings and threats may take the form of informing test-takers about the presence and effectiveness of detection techniques and the consequences of detection. Thus, it is not unreasonable to posit that informing job applicants that they will be retested under proctored conditions to verify their test scores may serve a deterrent function. Consonant with this, as previously noted, the **ITC Guidelines** (ITC, 2005) recommends informing test-takers in advance of the expectations and consequences of detected cheating. However, we again note that there is a paucity of empirical research investigating the effectiveness or efficacy of this technique in the context of unproctored Internet-based tests.

**Test Design Characteristics and Features**

One source of cheating arises from a breach of test security resulting in the preknowledge of test items. Thus, an unproctored Internet-based testing program that continuously and frequently administers the same test form (especially if its length is relatively short) is likely to be most susceptible to this cheating threat because of the high item overlap rate (the number of overlapping items encountered by test-takers divided by the length of the test; Chang & Zhang, 2002; Drasgow, Nye, Guo, & Tay, 2009). Strategies to make it more difficult for cheating to occur from this source include limiting the number of administrations, using multiple test forms, and using computerized adaptive tests (Foster, 2009). Thus, Burke (2009) describes “a randomized testing model through which equivalent but different tests are constructed from item response theory calibrated item pools”
So with this and other similar approaches, such as those in which they used an unproctored Internet-based speeded perceptual speed and accuracy test, Nye, Do, Drasgow, and Fine (2008) obtained results that were quite similar to Arthur, Glaze, Villado, and Taylor’s results. Specifically, the proctored retest scores were higher than the unproctored first test scores ($d = 0.22$; average retest interval = 1 month). In addition, on the basis of a regression-based approach, they identified only four individuals (out of 856) whose test score differences warranted some concern. However, it is worth noting that, whereas speededness may deter some forms of cheating, it does not curtail the use of surrogate test-takers or preknowledge of test items.

Finally, Foster (2009) describes an alternative test item format, the Foster Item, in which response options are presented serially instead of simultaneously and presentation ceases once an item has been answered either correctly or incorrectly. Although Foster does not describe it in these terms, it would seem that this item format shares several characteristics in common with the constructed-response format (for example, see Arthur, Edwards, & Barrett, 2002; and Edwards & Arthur, 2007). Concerning its effectiveness, Foster notes that “early research indicates that the Foster Item performs as well or better than its traditional multiple-choice counterpart with significant security advantages” (Foster, 2009, p. 32).

In summary, there are several approaches and techniques to deterring cheating. These techniques focus primarily on either discouraging cheating or making it more difficult to do so. However, as is characteristic of cheating on unproctored Internet-based tests in general, there is very limited empirical research that has investigated the efficacy and effectiveness of these approaches. So, in the absence of extensive empirical support, we only have their conceptual merit to go on—and said conceptual merit appears to be reasonably sound.

**Noncognitive Tests and Measures**

We use the terms “noncognitive” and “nonability” to broadly refer to the class of tests and measures for which there are ostensibly no true correct or incorrect answers to the items on the measure. These measures also typically entail self-reports.
Thus, although we may frequently make specific references to personality measures (because the extant literature has most frequently and extensively focused on this class of noncognitive measures), our discussion of specified issues is equally applicable to other self-report measures, such as measures of integrity, interests, attitudes, and other noncognitive constructs (Alliger & Dwight, 2000; Grubb & McDaniel, 2007; McFarland & Ryan, 2000). Furthermore, consonant with the construct/method distinction (Arthur & Villado, 2008), some testing methods have received some research attention in terms of dissimulation. These include resumes, job application blanks (for example, Wood, Schmidke, & Decker, 2007), employment interviews (for example, Delery & Kacmar, 1998; Ellis, West, Ryan, & DeShon, 2002; Levashina & Campion, 2006), biographical measures (for example, Schmitt & Kunce, 2002), and assessment centers (McFarland, Yun, Harold, Viera, & Moore, 2005).

So, in the absence of true correct or incorrect answers coupled with the inability to verify the accuracy of test-takers’ responses, noncognitive measures are recognized as being susceptible to test-takers’ dissimulation and response distortion which may take the form of self-deception or impression management efforts (Edens & Arthur, 2000). Consequently, whereas cheating is the malfeasant behavior of interest with ability and knowledge tests, dissimulation or response distortion—in the form of social desirability responding—is the primary malfeasant behavior of interest with noncognitive and nonability measures.

Paulhus (1986, 2002) highlights the distinction between self-deception and impression management as facets of social desirability responding. Social desirability responding is the tendency to over-report socially desirable personal characteristics and to under-report socially undesirable characteristics. It entails the tendency to choose specified responses even if they do not represent one’s true tendency or opinion. As a facet or dimension of social desirability responding, self-deception occurs when an individual unconsciously views himself or herself in an inaccurately favorable light; this typically entails a lack of self-awareness. In contrast, impression management or deliberate response distortion refers to a situation in which an individual consciously presents himself or herself falsely to create a favorable impression. Our focus here

is on intentional response distortion (that is, impression management), as opposed to self-deception.

A variety of terms and labels are used to describe response distortion in the extant literature. Some of these include social desirability, faking, dissimulation, impression management, lying, honesty, frankness, claiming unlikely virtues, denying common faults and unpopular attitudes, exaggerating personal strengths, response fabrication, good impression, and self-enhancement. Although there may be subtle distinctions among some of these descriptive labels, for the purposes of this chapter, we use the term “response distortion” to collectively refer to them and subsequently define it as a conscious and deliberate attempt on the part of test-takers to manipulate their responses in order to create an overly positive impression that deviates from their true standing on the trait or characteristic of interest (Ellingson, Sackett, & Connelly, 2007; McFarland & Ryan, 2000; Zickar & Robie, 1999). Response distortion is commonly conceptualized as systematic error variance (Arthur, Woehr, & Graziano, 2001; Tett, Anderson, Ho, Yang, Huang, & Hanvongse, 2006; cf. Uziel, 2010). Thus, job applicants are assumed to distort their responses because it assists them in attaining valued outcomes such as jobs and promotions. Hence, response distortion is posited to be determined by one’s motivation, and ability, along with specified situational factors such high- versus low-stakes testing (McFarland & Ryan, 2000, 2006; Snell, Sydell, & Lueke, 1999; Tett, Anderson, Ho, Yang, Huang, & Hanvongse, 2006).

Amount of Response Distortion

Paralleling and analogous to efforts taken to maintain test security and prevent cheating and other sorts of malfeasant behaviors in ability testing contexts, the prevailing view in both the academic and applied literatures is that applicants do distort their responses and answers on noncognitive and nonability self-report measures, although the focus has particularly been on personality measures (Anderson, Warner, & Specter, 1984; Birkeland, Manson, Kismore, Brannick, & Smith, 2006; Cellar, Miller, Doverspike, & Klawsky, 1996; Ellingson, Sackett, & Connelly, 2007; Hough, Eaton, Dunnette, Kamp, & McCloy,
expectation that malfeasant responding results in elevated scores on desirable characteristics and depressed scores on negative ones with said scores being inaccurate representations of the test-taker’s standing on the noncognitive constructs of interest. In addition, some empirical evidence indicates that test-takers who distort their responses are evenly distributed across the score range, with a slight trend of more malfeasant responders in the upper quartiles (Arthur, Glaze, Villado, & Taylor, 2010).

**Differences in Response Distortion in Proctored and Unproctored Settings**

The critical question is whether the unproctored remote delivery of noncognitive measures results in greater levels of response distortion compared to proctored testing. As previously noted, response distortion is posited to be determined by one’s motivation and ability, along with specified situational factors. Hence, proctored and unproctored settings could be conceptualized as differing on the extent to which they make it easy or difficult to fake. So, in an unproctored environment, access to illicit aids may create a relatively more permissive environment compared to proctored testing as test-takers may collaborate with other individuals (for example, surrogate test-takers or advisors) in an effort to inflate their test scores. However, it is unlikely that test-takers will engage in these behaviors if they are confident in their ability to elevate their test scores using their own personal schema of a desirable personality profile.

In response to concerns about response distortion, several techniques have been proposed for preventing or minimizing malfeasant responding on personality and other noncognitive measures. These include the use of forced-choice responses, empirical keying, warnings of verification, and response elaboration (for example, see Hough [1998] for a review). However, glaringly absent from this list of techniques is the use of test-proctors because the presence or absence of test-proctors has little or no bearing on controlling for response distortion. So, given the preponderance of research that indicates test-takers can effectively distort their responses under proctored conditions (Viswesvaran & Ones, 1999), there is little reason or impetus
for them to behave any differently under unproctored conditions (for example, use surrogate test-takers). Thus, the magnitude and extent of response distortion should be similar for both proctored and unproctored Internet-based noncognitive measures (Arthur, Glaze, Villado, & Taylor, 2010; Bartram & Brown, 2004; Gupta, 2007; Kaminski & Hemingway, 2009; Templar & Lange, 2008). In summary, in contrast to ability and knowledge tests, concerns regarding response distortion on noncognitive and nonability tests (for example, personality) are the same in both proctored and unproctored testing conditions.

Detection

Strategies or attempts to deal with response distortion generally take one of two forms—(1) detection and (2) deterrence. Detection techniques may take the form of score comparison and verification testing, the use of lie scales, inconsistency responding, response latencies, and statistical detection and control. Deterrence strategies include the use of forced-choice response formats, empirical keying, warnings, verifications, and threats, elaboration, and profile matching and the use of nonlinear models. These strategies and their efficacy and effectiveness are briefly reviewed below. It should be noted that the techniques and associated literature are based primarily on proctored tests. But as previously noted, because one would not expect differences in response distortion as a function of proctoring or lack thereof on noncognitive measures, these techniques (for example, use of lie scales) are equally applicable to and germane for unproctored testing as well. In addition, because of the technology via which they are delivered, there are some detection and deterrence techniques that are available for unproctored tests (for example, response latencies, and interactive prompts or cautions) that are unavailable for typical paper-and-pencil tests in proctored settings.

Score Comparison and Verification Testing

As with ability and knowledge tests and measures, one could conceivably use verification retesting and subsequent score comparisons as a means to detect response distortion on noncognitive measures. However, since proctoring is not a means by which one controls for response distortion, the expectation is that the levels of response distortion under unproctored and proctored conditions would be similar. To this end, although we were unable to locate any studies that undertook a within-study comparison of proctored and unproctored noncognitive measures, between-study comparisons of studies that have focused on either proctored or unproctored noncognitive measures provide some preliminary evidence. Thus, the results of both of Arthur, Glaze, Villado, and Taylor’s (2010) studies supported the supposition that unproctored noncognitive measures display levels of response distortion that are similar to those reported for proctored measures in the extant literature (for example, Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). So, on the basis of Arthur, Glaze, Villado, and Taylor’s (2009; 2010) data, one can reasonably conclude that unproctored personality measures display mean score shifts between high- and low-stakes testing conditions that are similar to those reported for proctored measures. In addition, the magnitude of the high-versus low-stakes score elevation and the percentage of individuals identified as having elevated high-stakes scores are also similar to those reported for proctored tests (Griffith, Chmielowski, & Yoshita, 2007).

In summary, because proctoring is not a means by which response distortion is controlled, verification retesting and subsequent score comparisons have different efficacy implications for cognitive versus noncognitive measures. Specifically, the effectiveness and meaningfulness of this detection technique with noncognitive measures is of very limited and questionable value and hence of limited utility as well. Consequently, it is not surprising that verification retesting and score comparison is not used or extensively discussed as a detection technique for noncognitive tests and measures.

Lie Scales

Because response distortion is theorized to be a consciously motivated behavior driven by a complex array of individual and situational factors, the propensity and tendency to engage in response distortion has been conceptualized as a discernable individual difference. Specifically, attitudes toward faking and
subjective norms predict intentions to fake which in turn, predicts faking behavior (McFarland & Ryan, 2006). Consonant with this, socially desirable responding has been demonstrated to differentiate individuals on their tendency to systematically describe themselves favorably (Paulhus, 2002; cf. Uziel, 2010) as well as their ability to fake (McFarland, & Ryan, 2000).

Concomitant with its conceptualization as an individual difference variable, the extant research also indicates that socially desirable responding can be reliably and validly measured and several approaches have been taken to accomplish this. The first entails a reliance on direct evidence of the tendency to engage in response distortion. An approach to obtaining this direct evidence is by means of a measure external to the focal personality measure—what we broadly refer to as “lie scales”. Examples of lie scales are the Unlikely Virtues Scale (Hough, 1998), and the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960).

A second approach is to embed the lie scale in the focal measure. With this approach, lie scale items are interlaced into the focal measure. Examples of personality measures that use this approach include the California Psychological Inventory (Gough & Bradley, 1996), the Hogan Personality Inventory (Hogan & Hogan, 1992), the Occupational Personality Questionnaire (SHL Group, 2000), the Personality Research Form (Jackson, 1999), and the Inwald Personality Inventory (Inwald, 1992; The interested reader is referred to Uziel [2010] for an informative article on the rethinking of social desirability scales and what they measure.)

Regardless of which approach is used, researchers and practitioners examine test-takers’ scores on the lie scales to determine whether the measure was answered honestly (Kuncel & Borneman, 2007). If a predetermined scale score is exceeded, then it is inferred that the test-taker may not have responded truthfully to the focal measure (for example, see Table 1 of Goffin and Christiansen [2003] for recommendations regarding what actions to take for high levels of social desirability responding as specified for several major personality inventories). Concerning the efficacy and effectiveness of lie scales in detecting response distortion, the research evidence indicates that inter-correlations among personality dimensions and lie scales increase under instructions to fake (for example, Michaelis &

Eysenck, 1971). Likewise, Stanush (1996) demonstrated that when test-takers are instructed to fake, there is a stronger relationship between lie scales and other personality inventory scales ($r = .34$), compared to test-takers instructed to answer truthfully ($r = .09$). Furthermore, instructions to fake resulted in elevated scores for both lies scales ($d = 0.82$) and scores on the Five Factor Model (FFM) personality factors ($d = 0.29, 0.53, 0.29, 0.36,$ and $0.28$ for agreeableness, conscientiousness, extraversion, emotional stability, and openness, respectively). The larger effects for elevated scores obtained for conscientiousness and emotional stability are consonant with more recent research as well (for example, Birkeland, Manson, Kisamore, Brannick, & Smith, 2006).

In summary, lie (social desirability) scales effectively detect intentional distortion (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). However, most, if not all, of this research has been conducted under proctored settings. Nevertheless, for reasons noted earlier, these relationships are expected to be directly applicable to unproctored settings as well; a conclusion that is consonant with data reported by Arthur, Glaze, Villado, and Taylor (2010).

**Inconsistency Responding**

An alternative approach to obtaining direct evidence of response distortion is to examine response patterns to indirectly assess response distortion. Indirect approaches have historically focused on detecting response styles, but may also be used to capture the presence of response sets. Response styles are tendencies to respond to items in specific ways regardless of their content (for example, acquiescence [agreeing with every statement]). In contrast, response sets are tendencies to respond to test content with a particular goal in mind (for example, answering in a manner that is socially desirable). Response styles are biases that are consistent across time and test content, whereas response sets are specific to a given situation. Controlling for response styles is usually done by including matched pairs of items with one item of a pair reversed. Inconsistent responses across item pairs is considered to be indicative of a response style (for example, agreeing to all items) or a response set (for example, agreeing to all items in the hopes that said responses are the most desirable). As an example, the Guilford-Zimmerman
Temperament Survey (Guilford, Zimmerman, & Guilford, 1976) uses this approach to assess the presence of response styles and sets.

In summary, inconsistency responding could be viewed as a variation of the lie scale approach to detect response distortion—albeit a more indirect one. Consequently, there is reason to expect that its efficacy and effectiveness at detecting response distortion would be similar to that observed for lie scales.

**Response Latencies**

It has frequently been shown that participants take longer to respond in fake good conditions compared to those responding under honest conditions (for example, Dwight & Alliger, 1997; Holden, 1995; Holden, Krone, Fekken, & Popham, 1992; Leonard, 1996; Robie, Brown, & Beatty, 2007; Vasilopoulos, Cucina, & McElreath, 2005; cf. Robie, Curtin, Foster, Phillips, Zylut, & Tetrick, 2000; Vasilopoulos, Reilly, & Leaman, 2000). The conceptual basis for this effect is that it takes longer to construct a distorted response than to answer honestly. And because unproctored Internet-based tests are by definition computer-administered, albeit via the Internet, assuming the test platform technology lends itself to it, then another possible means of detecting response distortion is to use response latencies. So, although the precision of the measurement of response times over the Internet would be a particularly important technological concern, response latencies as an indicator of response distortion would appear to be an avenue worthy of future research consideration.

**Statistical Detection and Control, and What Does One Do with Fakers?**

Regardless of the specific response distortion detection technique that is used, at some point a cutoff score has to be established above which test-takers are presumed to be intentionally distorting their responses and below which they are not. This is particularly true for the use of lie scales. Consequently, determining a cutoff score above which applicants are considered to be distorting their responses is an important step in detecting response distortion. Standardized tests and measures will have the specified cutoff scores reported in their test manuals. But regardless of whether one is using a standardized or an in-house test or measure, the issue of misclassifications—either in the form of false positives or false negatives—is critical. Hence, when deploying a cutoff score, it is important to consider the proportion of each type of misclassification that the specific score will produce. The mean plus three standard deviations has been used as a cutoff score (for example, Hough, 1998), although other values including the mean plus one or two standard deviations, and the mean plus 1 standard error of the measurement have been used as well (for example, Arthur, Glaze, Villado, & Taylor, 2010; Griffith, Chmielowski, & Yoshita, 2007; Hogan, Barrett, & Hogan, 2007).

Once detected, the question becomes what does one do with those who intentionally distort their responses? Two strategies may be used to reduce the effect of response distortion once it has been detected (Hough, 1998). The first entails removing applicants who are identified as distorting their responses from the applicant pool. Hough (1998) examined the effect of removing applicants who scored more than three standard deviations above the mean on an unlikely virtues scale and concluded that such a procedure reduced the effect of response distortion without affecting criterion-related validities.

The second strategy for addressing response distortion is to statistically adjust or correct the focal construct scores of either all applicants or only those identified as having distorted their responses based on their response distortion (for example, lie scale) score. The function of statistical control techniques is to statistically remove the irrelevant systematic error introduced by intentional response distortion from the focal construct scores. That is, statistical control techniques attempt to correct test scores so that they are not influenced by response distortion. To be effective, statistical control techniques require that specified relationships be present. Specifically, there must be (1) a relationship between the response distortion scale scores and the focal construct (for example, personality) scale scores; (2) a relationship between the focal construct scale scores and the job performance scale scores; and (3) an absence of a relationship between the response distortion scale scores and job performance. In the absence of these specific relationships, statistical control will be unable to accurately adjust test scores.
Although statistical control techniques have been examined using applicant samples (Hough, 1998; Ones, Viswesvaran, & Reiss, 1996), it is unclear whether these techniques have ever been employed in operational contexts. Nevertheless, as previously noted, studies that have examined statistical control techniques have concluded that they reduce the effects of response distortion without affecting criterion-related validities (Hough, 1998; Ones, Viswesvaran, & Reiss, 1996). In spite of these favorable results, statistically adjusting test scores based on lie scale scores as a means of controlling for response distortion remains controversial. Often, the necessary assumptions regarding the interrelationships between lie scale scores, personality scale scores, and job performance are tenuous at best. Consequently, support for the efficacy and effectiveness of this approach has generally not been very favorable and is mixed (Christiansen, Goffin, Johnson, & Rothstein, 1994; Ellingson, Smith, & Sackett, 2001; Goffin & Christiansen, 2003; Hough, 1998; Ones, Viswesvaran, & Reiss, 1996; Schmitt & Oswald, 2006).

A second concern with statistical control or score adjustments is that true variance may be removed from personality scale scores if response distortion is a true individual personality difference (Ones, Viswesvaran, & Reiss, 1996). A final concern questions the ability of statistical control techniques to correct responses that are intentionally distorted. As Cronbach (1990) wrote, “Once test users take a wrong course, there is no going back to the choice point” (p. 521). Phrased another way, if the test-taker tells lies to the tester, there is no way to convert the lies to truth.

In summary, based on the extant literature, the statistical control of response distortion does not appear to be strongly recommended and should probably be avoided. A number of researchers have concluded that correcting applicant test scores for socially desirable responding is neither a particularly effective nor viable approach to dealing with response distortion (see Hough & Furnham [2003] and Smith & Robie [2004] for summaries of this work). For instance, research such as Ellingson, Sackett, and Hough (1999), has shown that the detection and correction approach for handling response distortion does not work (that is, “corrected” scores are not accurate reflections of honest scores). More recently, Lönqvist, Paunonen, Tuulio-Henriksson, Lönqvist, and Verkasalo’s (2007) results indicate that (using uncorrected scores) even rank-order stability is maintained for applicants who are initially tested as applicants, and then tested three years later as incumbents (cf. Arthur, Glaze, Villado, & Taylor, 2010).

**Deterrence**

As with ability and knowledge tests, the methods and techniques that have been developed and used to deter response distortion can take several forms. For instance, features such as forced-choice responding (all response choices are designed to be similar in terms of social desirability) and empirical keying to create subtle indicators of the focal construct can be built into the test during its design and construction to minimize and deter response distortion (Hough, 1998). The use of firm test instructions and warnings have also been investigated as a means of deterring and minimizing response distortion (McFarland & Ryan, 2000; Vasilopoulos, Cucina, & McElreath, 2005). These strategies, which vary in their degree of effectiveness at deterring response distortion, are briefly reviewed below.

**Forced-Choice Response Formats**

Forced-choice response formats entail the use of equally desirable response options or items and force the test-taker to choose between them. In spite of their intended goal of reducing response distortion, the research evidence suggests that forced-choice strategies are not immune to intentional distortion (Hough, 1998). Waters (1965), in a review of forced-choice inventories, concluded that individuals are able to distort their responses. And so although it is posited that forced-choice scales may reduce response distortion (Hirsh & Peterson, 2008; Jackson, Wrobleski, & Ashton, 2000; Snell, Sydell, & Lueke, 1999), they do not eliminate response distortion (Christiansen, Burns, & Montgomery, 2005; Converse, Oswald, Imus, Hedricks, Roy, & Butera, 2008; Heggestad, Morrison, Reeve, & McCloy, 2006). Furthermore, ipsative scales make comparisons between applicants difficult (Hough, 1998; Heggestad, Morrison, Reeve, & McCloy, 2006; McCloy, Heggestad, & Reeve, 2005).
Empirical Keying

Empirical keying, which is commonly associated with biodata inventories, is a scoring procedure that focuses on the prediction of an external criterion using keying procedures at either the level of the item or item-option. In using empirical keys, the valence of specified items and item-options is less transparent and permits the design of a more subtle assessment of the non-cognitive constructs of interest. Consequently, subtle scales are posited to be less susceptible to response distortion than obvious, more transparent scales. However, although they may minimize it, empirical keying does not eliminate response distortion since research suggests that empirically keyed measures are still susceptible to response distortion (Hough, 1998; Kluger, Reilly, & Russell, 1991; Visweesvaran & Ones, 1999).

Warnings, Verification, and Threats

Warnings, verification, and threats take the form of informing test-takers that their answers will or can be verified (McFarland, 2003; McFarland & Ryan, 2000; Paco & Borman, 2006; Vasilopoulos, Cucina, & McElreath, 2005). Assessments of the effectiveness of this deterrence strategy have yielded mixed results (Converse, Oswald, Imus, Hedricks, Roy, & Butera, 2008; Dwight & Donovan, 2003; Hough, 1998) because of the possible unintended consequences resulting from their use (Robson, Jones, & Abraham, 2008; Vasilopoulos, Cucina, & McElreath, 2005). For instance, Vasilopoulos, Cucina, and McElreath found that a warning strategy had the unwanted consequence of increasing the complexity of the personality measure. They found that a personality measure preceded by a warning was correlated with a measure of cognitive ability and concluded that the warnings of verification made responding so complex that the personality measure was measuring cognitive ability to a limited degree.

Interactive Prompts or Cautions

Because of the technology with which they are administered, the inconsistency responding approach previously described can be extended into a deterrence strategy. Specifically, pairs of items for which one would expect test-takers to respond consistently can be designed into the test and whenever a test-taker responds to the second item of a pair inconsistently, the system could be programmed to give some sort of interactive prompt to encourage or caution the test-taker to pay attention to the item content and respond accordingly. Although this approach is novel, it can to some extent be considered to be a variation of the warnings approach discussed above. It is also worth noting that distorting consistently to both elements of an item pair makes it difficult, if not impossible, to detect distortion using this approach.

Elaboration

Elaboration is a strategy whereby test-takers are asked to elaborate on their responses. For example, Schmitt and Runce (2002) asked applicants to provide an elaboration only if they had endorsed specific responses (sometimes, often, or very frequently) to a question (such as, How often have you rearranged files [business, computer, personal] to make them more efficient in the past year?). The efficacy of this strategy is based on the premise that by asking test-takers to elaborate on and provide detailed follow-up responses to their answers, they are less likely to distort their responses because they would also have to concoct an elaboration as well. Although elaboration strategies appear to be somewhat effective, these strategies may also introduce unwanted consequences. For example, Schmitt, Oswald, Kim, Gillespie, Ramsay, and Yoo (2003) found that requesting respondents to elaborate only if they endorsed specific responses resulted in lower overall test scores. Thus, it appears that elaboration may discourage all applicants from endorsing responses that require additional work. However, Schmitt, Oswald, Kim, Gillespie, Ramsay, and Yoo report that the correlation between test scores and performance were equivalent across elaboration and no elaboration conditions.

Profile Matching and Nonlinear Models

Profile matching entails the matching of a pattern of applicant scores across multiple dimensions or constructs to some ideal or standard profile. Thus, the use of profile matching or profile similarity indices (which are used extensively with measures such as the Guilford-Zimmerman Temperament Survey [GZTS; Guilford, Guilford, & Zimmerman, 1978]), are
an attempt tocompare two sets of multiple personality dimensions (for example, profiles) representing, for example, an applicant and an “ideal” employee, via a single score or index that provides information on the degree of congruence, similarity, or match between the two profiles. Profile similarity indices used in congruence research can be classified into one of two categories—those representing the correlation between the two profiles and those based on the sum of differences between profile elements (personality variables or dimensions; Edwards, 1993). Edwards (1993) presents a detailed description and review of specific indices of these two types of profile similarity indices along with a discussion of methodological problems associated with their use in congruence research, including discarding information regarding the absolute level of the profiles, along with the direction of their difference, and with correlations, the magnitude of the difference as well. He also notes that profile similarity indices mask which elements are responsible for the differences between the profiles. Given these methodological problems, Edwards recommends polynomial regression procedures and shows how they may be used to avoid the problems with profile similarity indices while capturing the underlying relationships profile similarity indices are intended to represent. (The reader is referred to Edwards for a more in-depth, detailed coverage of these issues. Also see Kristof [1996] for additional discussion of these issues and some limitations associated with polynomial regression analysis.)

Inherent in the use of profile matching is an implicit, if not explicit, recognition of the nonlinearity of the specified relationships. The issue of response distortion in the extant literature and its associated discussion therein is predicated on the assumption that the test scores are being used in a linear fashion such that higher scores on the focal constructs are generally deemed to be linearly better (cf. Arthur, Glaze, Villado, & Taylor, 2010). However, one can envisage several conceptually and theoretically sound scenarios whereby the relationship between personality variables and job performance is better conceptualized as being nonlinear. For instance, one can envisage a situation in which moderate levels of agreeableness may be related to effectiveness in customer relations, with low and high levels of agreeableness, on the other hand, being somewhat counter-productive (Graziano & Eisenberg, 1997; Graziano, Jensen-Campbell, & Hair, 1996). Likewise, it is possible to be too conscientious to perform certain roles effectively as is reflected in the obsessive-compulsive label. Murphy (1996) comments on this possibility when he notes that an individual who is high on conscientiousness “might be so conventional and rule-bound that he or she cannot function in anything but the most bureaucratic setting” (p. 22).

Accordingly, the relationships among various personality constructs and job performance may be better conceptualized, under some circumstances, as being nonlinear. Thus, assuming there is empirical support for such an approach (see Day & Silverman, 1989; Robie & Ryan, 1998; Robins, 1995; Scarborough, 1996; Sinclair & Lyne, 1997 as examples of studies that have explored nonlinearity in the relationships between personality variables and job performance), profile matching, coupled with its underlying use of nonlinear models, may mitigate concerns about response distortion, specifically, the uniform elevation of scores. So, to the extent that the specific profile or score configuration is unknown to test-takers, coupled with the fact that the ideal profile or score configuration is usually organizationally job-specific, one would expect profile matching approaches to be less susceptible to the ubiquitous effects of response distortion, especially that which takes the form of across the board elevation of scores, particularly on dimensions that are fairly transparent. In summary, although we were unable to locate any empirical research that investigated the efficacy and effectiveness of the use of nonlinear models as a deterrent technique, response distortion may be less of an issue or concern with this approach because dissimulation to obtain high construct-level scores does not necessarily imply or result in the successful faking of the optimal profile. So, because scores are not being used in a linear fashion, response distortion may be less of a concern.

**Conclusion**

It is widely accepted that the remote delivery of assessments in the context of personnel selection and related employment decision-making is increasingly becoming a common practice
sources of cheating must be curtailed using alternative methods and technologies. As a summary statement, the efficacy and effectiveness of these alternative methods and technologies are not well understood. For example, the use of webcams as a means of electronic monitoring may be intuitively appealing, but in the absence of empirical data, their efficacy remains unknown. Furthermore, it is unclear how feasible the use of some of these technologies are for unproctored testing (for example, iris and retina scans). That is, although it may be reasonable to expect test-takers to have access to webcams, it may be unreasonable to expect them to have access to biometric identification equipment. In the absence of bona fide detection techniques, the use of proctored retesting is the only viable method of detecting cheating. However, proctored retesting reduces the cost-effectiveness and other advantages of unproctored testing (test anywhere at any time). Furthermore, there is currently no consensus on the extent to which unproctored and proctored test scores must diverge to raise concerns and warrant corrective action.

Regardless of the specific methods or technologies used to detect it, the issue of taking corrective action when cheating is detected is a complex one. For instance, one must first determine what constitutes evidence of cheating. The academic and educational testing literature reflects concerns about accusing or confronting test-takers about cheating (Cizek, 1999). Furthermore, professional and ethical guidelines require test-takers to be informed of testing irregularities, and the implications of test-takers reactions to being accused of cheating has not been investigated. On a related note, there is a dearth of information regarding the legal implications of canceling or correcting applicants’ test scores because of suspected cheating.

Given the complexity and difficulty of detecting cheating and compiling convincing evidence that cheating has occurred, attempting to deter cheating may be more efficacious. Approaches to deterrence can take one of two forms: (1) increasing the saliency of detection methods and (2) using test design characteristics that make it more difficult to cheat. Increasing the saliency of detection techniques can take the form of overt monitoring and warnings and threats. Although the efficacy of these methods is unknown, it is reasonable to posit that warnings
(especially of proctored retesting) may curtail cheating attempts. Regarding test design characteristics, the use of speeded tests (Arthur, Glaze, Villado, & Taylor, 2010), measuring constructs that are difficult to cheat on (Nye, Do, Drasgow, & Fine, 2008), limiting item exposure (Drasgow, Nye, Guo, & Tay, 2009), and guarding against preknowledge of test content would seem to be effective strategies. However, these methods engender some disadvantages. Using speeded tests is limited to situations in which a speeded test is consonant with the job-relatedness of the test, and only a limited number of constructs (for example, perceptual speed) may be particularly resistant to some forms of cheating. In addition, limiting item exposure and preknowledge of test content may be administratively difficult.

So, as a summary statement, it is reasonable to assume that cheating will be quite widespread on unproctored ability and knowledge tests if left unchecked. Methods for curtailing cheating include deterring and detecting cheating behaviors. However, most of the methods and techniques for curtailt theying cheating behavior have limited or no empirical support. Furthermore, there are theoretical, practical, ethical, and legal issues that are yet unresolved regarding deterring and detecting cheating, the evidence required to take corrective action, and the proper corrective action to take.

For noncognitive and nonability tests and measures, response distortion on unproctored Internet-based tests seems to be no more pervasive or extensive than that for proctored testing since proctoring is not a method for curtailing response distortion. Concerning methods for detecting and deterring response distortion, by virtue of the technology by which they are administered, there are methods that are unique to unproctored Internet-based tests, and those that are common or applicable to both proctored and unproctored Internet-based tests. The latter include the use of forced-choice response formats, empirical keying, warnings, verifications, and threats, elaborations, lie scales, and indicators of inconsistency responding. Response distortion detection and deterrence approaches that are unique to unproctored Internet-based tests include the use of response latencies and interactive prompts or cautions. Compared to the literature on cheating on unproctored ability and knowledge tests, there is relatively more empirical literature investigating the effectiveness of these methods, but most, if not all, of it has been conducted in the context of proctored tests. However, as previously noted, there is no conceptual reason to suspect that these findings would not generalize to unproctored settings. Finally, as with cheating on ability and knowledge tests, there are unresolved issues regarding deterring and detecting response distortion on noncognitive measures, the evidence required to take corrective action, and the proper corrective action to take.

In addition to the preceding summary statements, there are a number of noteworthy observations. First, it is important to note that the proctored versus unproctored delivery of assessments has more profound effects on and implications for ability and knowledge tests than it does for noncognitive and nonability measures primarily because proctoring is not a means for controlling for response distortion on the latter type of assessments. As a result, for noncognitive tests and measures, response distortion issues are common across both unproctored and proctored settings (Arthur, Glaze, Villado, & Taylor, 2010; Bartram & Brown, 2004; Gupta, 2007; Kaminski & Hemingway, 2009; Templer & Lange, 2008). However, because of its technological mode of administration, there are some issues, such as the use of response latencies for detection and interactive prompts as a deterrent, that may be more commonly associated with the remote delivery of noncognitive and nonability measures than they are with ability and knowledge tests.

Second, although we reviewed and discussed them in a singular fashion, for most, if not all, the detection and deterrence techniques for both ability and noncognitive measures, multiple methods can be used in conjunction. That is, the use of these methods and techniques is not mutually exclusive. So, for instance, for noncognitive measures, response latencies may be used in addition to lie scales, forced-choice response formats, and empirical keying. Consequently, research investigating the efficacy and effectiveness of various combinations of techniques would be informative. However, even in the absence of said research, it does not seem unreasonable to posit that the conjunct use of
various techniques could be more effective than the use of any one method by itself.

Third, the amount of research in this domain is very limited and is definitely not commensurate with the use of these types of assessments in practice. As a result of this, we were unable to directly comment on or evaluate the effectiveness and efficacy of several of the detection and deterrence methods such as the use of verification retesting (as a deterrent) and webcams (as a detection method). Future research should investigate, for example, the extent to which verification retesting deters cheating on initial testing and applicant reactions to this method. As previously mentioned, the use of verification retesting implicitly undermines the perceived veracity of initial unproctored test scores. Furthermore, in the absence of data to the contrary, it is not unreasonable to posit that requiring test-takers to own or obtain webcams may reduce the number of unproctored test-takers. This problem would be exacerbated if ownership of webcams varied as a function of status on a protected class variable.

Fourth, the distributional placement (position in the score distribution) of malevolent responders (cheaters and those who distort their responses) is a critical issue. Arthur, Glaze, Villado, and Taylor (2010) argue that the impact of malevolent responding on employment decisions is partially a function of distributional placement. That is, if the preponderance of malevolent responders is in the low end of the distribution, malevolent responding may have minimal effects. However, if malevolent responders are evenly distributed across the score range, or reside in the upper end of the distribution, then cheating and response distortion become relatively more critical issues. Similarly, Arthur, Glaze, Villado, and Taylor provide data that suggests that those suspected of cheating on a cognitive ability test were evenly distributed across the score range, with a slight trend of having more cheaters in the upper end of the distribution. In contrast, Impara, Kingsbury, Maynes, and Fitzgerald (2005) found that cheating occurs across the score range with the exception of those at the high end of the distribution. In the context of nonability tests, Arthur, Glaze, Villado, and Taylor found that response distortion occurred across the score range, with a slight trend towards having more response distortion in the upper end of the score distribution. Griffith, Chmielowski, and Yoshita (2007) examined the proportion of applicants who were suspected of distorting their responses that would be hired under varying selection ratios. Their results indicated that with a 50 percent selection ratio, 31 percent of the test-takers in the hiring range would not have been hired using their honest scores; this number increased to 66 percent with a 10 percent selection ratio. Similarly, inflating self-report SAT scores appears to be negligible for test-takers in the upper and midrange of SAT scores, however those in the lower end of the distribution tended to inflate their SAT scores (Newman & Lyon, 2009).

Fifth, there are a number of less commonly used approaches for detecting response distortion that are nevertheless, worth noting. For example, Zickar and Drasgow (1996) present an item-response theory-based theta-shift model for detecting response distortion. Specifically, they proposed that test-takers who distort their responses respond to a subset of items honestly but respond to some items in an inaccurately favorable manner. Thus, test-takers whose response patterns reflect a theta-shift are suspected of response distortion. As a summary statement, the theta-shift model resulted in a lower number of false positives compared to a social desirability scale. Zickar and Robie (1999) provided convergent evidence for the theta-shift model by examining response distortion at both the item- and scale-level of analysis. Their findings suggest that for test-takers with the same level of the underlying construct, those who distorted their responses were more likely to endorse a more extreme positive response. Both of these studies demonstrate the viability of item-response theory-based methods for detecting response distortion.

In conclusion, it is widely recognized that merit-based public-sector selection and promotion testing, especially municipal safety forces (police and fire), is extremely litigious. Thus, one could say that unproctored Internet-based testing has really “arrived” and overcome its security threats and concerns when it is widely used and accepted in this type of testing environment and setting. This would currently appear to be the case for biodata and training and experience measures, but less so, if not virtually nonexistent, for ability and knowledge-based assessments.
References


Industrial and Organizational Psychology: Perspectives on Science and Practice. (2009), 2(1).


Chapter Five

COMPUTERIZED ADAPTIVE TESTING
Rodney A. McCloy and Robert E. Gibby

Dating from the 1960s, computerized adaptive tests, or CATs, have been developed to meet a host of assessment needs. The basic premise behind a CAT is that it provides a test examiner or administrator the ability to individually assess a respondent by selecting and presenting items based on the respondent's ability or trait level (theta). Unlike a respondent's true score in classical test theory (Allen & Yen, 2002; Crocker & Algina, 1986), which is conditional on the test (or set of items) in question, the theta (or ability) score is a characteristic of the respondent and is independent of test content (Lord & Novick, 1968). Typically starting with an item around the middle of the theta distribution for the construct being assessed (an item of moderate difficulty), a CAT chooses and presents an item from a pool of items that has been calibrated against an item response theory (IRT; De Ayala, 2009; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Van der Linden & Hambleton, 1997) model that describes response behavior. The respondent then answers the item, and the CAT estimates a new provisional theta for the respondent based on the information provided by the response to this most recently administered item. Given this new estimated theta, the adaptive algorithm then goes to the item pool and selects the next item to present. All things being equal, the next item to