Ethical issues in the use of computerized assessment

Stefan E. Schulenberg"*-*, Barbara A. Yutrzenka`

"The University of Mississippi, PO Box 1848, University, MS 38677 1848, USA
`The University of South Dakota, 414 E. dark Street, Vermillion, SD 57069, USA

Abstract

Computer applications in the field of psychological test administration have significant ethical implications for clinicians, client/responders, and computerized test construction and administration. Lack of awareness of computer-related issues may undermine clinicians' ability to ethically perform computerized psychological assessments. Graduate training in computerized testing is limited, and clinicians should be exposed to ethical concerns, potential judgment errors, and possible pitfalls in evaluating computer-generated reports. Recommendations for clinical research and practice are offered. Computerization clearly presents a series of dilemmas for psychologists conducting clinical assessments that will continue well into the future. Increased awareness of relevant issues will enhance the chances that ethical dilemmas will be successfully navigated.

© 2003 Elsevier Ltd. All rights reserved.

Keywords: Ethics: Psychological assessment; Computer-assisted testing; Computer-assisted diagnosis; Computer attitudes

Mirroring the infusion of technology in nearly every facet of personal and professional life in contemporary society is the increasing presence of computers in clinical practice. Advancements in computer technology and the continued integration of this technology into clinical practice are changing the way clinicians conduct psychological assessments. For example, computers are being used for test administration, item scoring, transformation of scores, test interpretation, and data storage, as well as for rendering clinical judgments and report writing (Butcher, 2003; Garb, 2000a; Maddux & Johnson, 1998; Snyder, 2000; Wilkins & MacKenzie, 1989).

There is no doubt that computers have changed and will continue to change the way clinicians develop, administer, and use psychological assessment (Garb, 2000a,
With this certainty comes the concomitant need to examine the implications of this change, particularly the ethical issues involved in the use of computers in assessment. The ethical standards promulgated by the profession of psychology address ethical issues relating to assessment in general. Table 1 presents an overview of principles specifically relating to computerized assessment as presented in the *Standards for Educational and Psychological Testing* (1999; joint Standards authored by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education; hereafter referred to as the AERA Standards), the American Psychological Association’s (APA) *Ethical Principles of Psychologists and Code of Conduct* (2002), and the Association of State and Provincial Psychology Board’s ASPPB *Code of Conduct* (2001). However, these standards vary in the degree of specificity with which they address issues relating to computerized assessment. Thus, the purpose of this paper is to provide an overview of key ethical domains involved in computerized assessment. The domains reviewed relate specifically to competence, utilization of computerized test interpretation packages, accommodating the responder, equivalency between traditional and computerized assessment procedures, and confidentiality. Recommendations

<table>
<thead>
<tr>
<th>Ethical domain</th>
<th>Core issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence (AERA 12.1, APA Principle A, 2.01, 2.03, 2.04, 9.02, 9.06, 9.09; ASPPB A.1, A.2, A.4)</td>
<td>Clinicians should practice within their area of competence. They should be aware of the issues of incorporating technology into clinical practice. Clinicians are responsible for the applicability of computer-generated reports.</td>
</tr>
<tr>
<td>Computer-generated Interpretive Reports (AERA 5.11, 11.21, 12.15; APA 9.02, 9.06, 9.09; ASPPB H.3)</td>
<td>Clinicians should take into account the needs of the client. Responders should also be provided with information about the pros and cons of computerized assessment measures.</td>
</tr>
<tr>
<td>Accommodating Responders (AERA 5.11, 8.2, 8.3, 11.21, 12.15; APA 3.10, 9.02, 9.06; ASPPB C.1, H.3)</td>
<td>Documentation should be provided as to the interchangeability of scores across modalities. Publishers of computer-interpretive software should summarize the support for the results. Furthermore, documentation of the methods for scoring and/or classifying should be provided (including variance not related to the construct of interest, such as might be the case with a responder's familiarity with computers). Clinicians should solicit this information from testing companies if it is not readily available.</td>
</tr>
<tr>
<td>Using Scores Across Modalities-Equivalence (AERA 6.11, 6.12, 13.18; APA 9.02, 9.05, 9.06, 9.09; ASPPB H.5)</td>
<td>Clinicians should ensure the privacy of the responder.</td>
</tr>
<tr>
<td>Confidentiality (AERA 6.21; APA 4.01, 6.02; ASPPB E.6)</td>
<td></td>
</tr>
</tbody>
</table>

for research and clinical practice are offered following a discussion of these ethical domains.

1. Competence

An underlying assumption of professional practice is that practitioners practice within the boundaries of their particular areas of competence. As this relates to psychological assessment, clinicians are expected to be knowledgeable about and skilled with the administration and interpretation of the assessment techniques, tools, and measures they utilize (APA, 2002). Assessment competence, thus, includes familiarity with the research support and psychometric strengths and limitations of the assessment tools and procedures, accurate interpretation and presentation of assessment results, protecting the integrity of the assessment process through scientifically based test construction, limiting assessment to those with appropriate qualification, and protecting client confidentiality. Expectations of competence would logically extend to all practitioners involved in traditional as well as computerized assessment techniques, tools, and procedures. However, it is important to highlight that practitioners who are unfamiliar with the literature relating to the use of computer technology in clinical practice, who are uninformed about the psychometrics or appropriate methods of administering, scoring, or interpreting computerized tests, or who have limited knowledge about the general operation of the computer or limited computer skills, may be practicing outside their area of competence.

2. Computer-generated interpretive reports

The potential for clinicians to misuse computer test interpretations has been well documented (e.g. Maddux & Johnson, 1998; Matarazzo, 1986). Maddux and Johnson noted that it is currently inappropriate to use the majority of interpretation software because these programs "encourage the use of assessment instruments by personnel who are not fully competent in their use, and they apply a simplistic paradigm (IF-THEN) to the solution of complex human assessment problems." (p. 99) This potential simplification of the interpretation process opens the doors for unqualified practitioners to administer psychological tests and rely on the computerized interpretation to form their diagnostic impressions. This, in turn, may result in clinicians facing the possibility of litigation because of this misuse of the assessment process (e.g. Butcher, 1987; Butcher, Perry, & Atlis, 2000; Matarazzo, 1983, 1986).

Adding to the risk of misuse of computer test interpretations is the fact that the validity for many computerized test interpretation packages has not been definitively established, and some may be biased (Garb, 2000a). Garb noted that some interpretive packages may produce reports that are more accurate for certain racial/ethnic groups. Overall, validity studies on test interpretation software are relatively rare, and are totally nonexistent for most programs (Snyder, 2000). Similarly,
Butcher (2003) noted that some computer-based personality interpretive programs are published commercially with limited determined validity. Essentially, it falls to the clinician to ensure that the computerized interpretation is an accurate reflection of the respondent being assessed, and not just a series of interpretations put forth in a software package by a testing company (Butcher, 2003). Snyder (2000) notes that use of computerized interpretive reports is risky, and that clinicians may be liable if they include a computer-generated interpretation in a report without adequately evaluating its relevance.Clinicians should not depend on these interpretations unless they have the knowledge to determine their quality and their applicability to the responder (e.g. empirical support, limitations). Discrepancies between a computerized report and client characteristics need to be accounted for by the clinician (Butcher, 2003). Finally, computerized interpretive reports should be used as an adjunct to the clinical assessment (Butcher, 2003), and computers should only be used to assist in making final clinical decisions if they have been demonstrated to be more valid than clinicians' judgments (Garb, 2000a).

3. Accommodating the client/responder

In addition to being aware of issues relating to computerized test interpretation, clinicians also need to be aware of the array of potentially confounding responder variables that may differentially influence the process and outcome of a computerized testing session. Responders bring various predilections with them into computerized testing sessions and can have a variety of experiences during a testing session that can negatively influence reliability and validity. Two specific areas that can affect the administration of a computerized test are a responder's attitudes toward computers and their feelings of aversion toward computers.

3.1. Attitudes towards computers

A responder's attitudes toward technology may play a crucial role during computerized psychological test administrations. Several authors have reported that individuals completing computerized interviews or evaluations tend to be accepting of these procedures and in some cases may be more willing to provide sensitive, personal information to the computer (Dignon, 1996; Locke & Gilbert, 1995; Lucas, Mullin, Luna, & McInroy, 1977; Maddux & Johnson, 1998; Nurius, 1990; Olson, 2001; Petrie & Abell, 1994). Thus, on some occasions computers may lead to greater openness on the part of the responder. However, support for the computer's ability to solicit sensitive information or a greater degree of acceptability is not uniform (Erdman & Foster, 1986; Erdman. Greist, Gustafson, Taves, & Klein, 1987; Locke & Gilbert, 1995; Skinner & Allen, 1983). Spinhoven, Labbe, and Rombouts (1993) found that patients with less computer familiarity and negative computer attitudes did not derive as much satisfaction and were not as relaxed as those with more positive attitudes and greater computer familiarity. Clearly, attitudes toward technology have significant potential to influence interactions with computers.
3.2. Computer aversion/anxiety

Computer aversion, sometimes referred to as computer anxiety, may hinder a person's performance during a computerized test administration (Gardner, Discenza, & Dukes, 1993; George, Lankford, & Wilson, 1992; Hofer, 1985; Peterson, Johannsson, & Carlsson, 1996; Schulenberg & Yutrzenka, 1999, 2001). For instance, Schulenberg and Yutrzenka (1999) indicated that those who experience discomfort when using computers may have higher scores on computerized measures of negative affect than those without such discomfort. A user's discomfort with the technology could be a differentiating factor when taking a test on a computer as opposed to more traditional methods (Butcher, 2003).

The association between test anxiety and computer anxiety may also be a relevant issue, and one of continuing interest (e.g. Powers, 2001; Shermis & Lombard, 1998; Shermis, Mzumara, & Bublitz, 2001). Despite these recent papers, the issue remains as to whether computer anxiety is distinct from test anxiety, or whether it is part of it. This is an important concern as the relationship between the two forms of anxiety may significantly influence a computerized assessment. If a person experiences test anxiety but is generally comfortable and familiar with computers, would taking a test on a computer lessen test anxiety and enhance performance? How much measurement error would be introduced by the computer if someone who is anxious about testing and uncomfortable using computers is administered a computer administered examination as opposed to a paper and pencil instrument? Ultimately, clinicians who do not make adequate accommodations for persons who experience computer aversion risk inaccuracies in their interpretation of the test results.

4. Equivalence of scores across modalities

Computerized testing improves the reliability, standardization, and objectivity of test administration by administering items the same way each time (Butcher, 1987; Dignon, 1996; Kobak, Greist, Jefferson, & Katzelnick, 1996; Maddux & Johnson, 1998; Sturges, 1998). Computers also decrease scoring errors and increase processing capabilities such as being able to arithmetically or statistically transform scores (Butcher, Keller, & Bacon, 1985; Dignon, 1996; Kobak, Reynolds, & Greist, 1993; Maddux & Johnson, 1998; Sturges, 1998). Computerized testing also allows clinicians to collect other types of data, such as recordings of elapsed time between the presentation of an item and a response (referred to as response latency). However, despite the positive contribution of computerized assessment to clinical practice, there continue to be a number of issues related to the construction and administration of computerized tests that must be considered. For instance, there are currently no well-established theoretical foundations that clinicians may follow to generate psychological tests based on new technology, such as an interactive video assessment (Olson-Buchanan & Drasgow, 1999). The alternative to generating new measures specifically for computers is to adapt conventional measures into computerized formats. While this is becoming a routine part of clinical research and practice, test
construction issues such as equivalence must be considered (Schulenberg & Yutrzenka, 1999). That is, computerized tests must be empirically validated to determine their level of equivalence to their traditional counterparts, indicating that they produce basically the same results (e.g. Butcher, 1987, 2003; Butcher et al., 2000; Maddux & Johnson, 1998).

Many studies have examined the equivalence between traditional and computerized forms of the same instrument. These studies often support that conventional and computerized instruments are essentially equivalent. Maddux and Johnson (1998) noted "that when careful comparisons are made between paper-and-pencil and computer versions of certain types of tests, their equivalency is remarkably close." (p. 95). Support for the equivalence of instruments across modalities is not uniformly reported, however, and may be dependent on the domain of the test (e.g. personality test, neuropsychological test, interview). Butcher et al. (2000), reiterated in Butcher (2003), noted that the "comparability of computerized and standard administration of various measures appears to vary, with the most promising results in the area of personality assessment." (p. 9) Snyder (2000) reported that the equivalence literature in the area of personality assessment is "mixed" (p. 53), with the established equivalence outside the area of personality assessment being "weak or nonexistent" (p. 53). Thus, the equivalence data may not be as clear-cut as it may at first seem.

The issue of equivalence across paper-and-pencil and computer modes of administration becomes more complex when one considers measures of negative affect, primarily because there is evidence to suggest that factors such as computer aversion may act to artificially inflate scores during computerized assessments of negative affect (Schulenberg & Yutrzenka, 1999). A responder's "discomfort with computers and consequent awkwardness when dealing with them" could be a major factor in the lack of equivalence of results across paper-and-pencil and computer modalities (Butcher et al., 2000, p. 7).

Clinicians should consider the empirical basis for equivalence when using a different form of a particular instrument. For instance, does the computer version of a particular test demonstrate construct equivalence? Construct equivalence is important because computerization may alter a psychological test to the point that it may not be measuring the same construct as its traditional counterpart. Others have noted the need to examine construct validity (e.g. King & Miles, 1995; Neuman & Baydoun, 1998; Turban, Sanders, Francis, & Osburn, 1989).

5. Confidentiality

Finally, with increased use of computer technology in clinical practice, clinicians may unknowingly violate ethical standards relating to confidentiality, as would be the case if access to computer test results were not protected (Maddux & Johnson, 1998; McMinn, Buchanan, Ellens, & Ryan, 1999). Security of computer-generated materials is a responsibility of users of computer-based reports (Butcher, 2003). Storing confidential client information on the hard drive or on a network may
compromise confidentiality in certain situations, as may be the case when the computer requires service from a technician.

6. Recommendations for practice

According to AERA Standard 5.5, the clinician needs to ensure that responders are knowledgeable of the computer equipment and the response format, and that they are able to make adequate responses using the necessary equipment. Clinicians may be tempted to leave responders alone with the computer for extended periods of time, which may detract from the reliability and validity of the testing session as well as limit valuable observational data regarding the respondent's approach to test-taking and comfort level. Given the potential for computer aversion and attitudes toward computers to influence a responder's performance on a computerized measure, it is recommended that clinicians systematically assess for each construct as a routine part of their clinical interview. Careful probing and behavioral observations appear to be solid ways for clinicians to assess responders' level of comfort with computer technology.

There are several additional recommendations that may assist clinicians who use computer assessment in their practice. These include allowing responders the opportunity to decline a computerized test administration, ensuring the confidentiality of computerized assessment materials, becoming critical consumers of computer-based interpretive programs, knowing the psychometrics of the computerized assessment measure, and monitoring the qualifications of computerized assessment users.

6.1. Give responders the opportunity to decline (Dignon, 1996)

The choice whether a computer is introduced into a psychological assessment should be at the responder's discretion. Furthermore, commensurate with AERA Standard 8.2, responders should be informed of the pros and cons of taking a test in a computerized format, as opposed to a paper-and-pencil format. For instance, how easy is it for responders to change their answers in a computer-administered format? Also, despite the fact that there is a proliferation of studies available in the literature relating to computerized measures, clinicians need to be aware that more research tends to be available on paper-and-pencil measures. This may be of particular concern when assessing persons from populations less equally represented in an instrument's norm sample.

6.2. Ensure confidentiality

Computers containing confidential clinical information (such as test results, progress notes, reports, etc.) should be stored in a secure area whenever possible. Whether or not computers are stored in a secure area, access to clinical files should be restricted to individuals with an appropriate access code. Using numbers for clients
instead of their names can minimize the presence of identifying information in the system. Given that there may be variations from operating system to operating system, clinicians should also pursue expert advice as to the pros and cons of certain storage techniques. For instance, clinical files may be compromised if they are stored on the hard drive of a computer or on a network. In the case that a computer requires service, it may be useful to have a standing relationship with a specific company or person. A confidentiality agreement between the clinician's organization and the agency/technician would aid in protecting the privacy of responders should a technician come into contact with identifying data during the course of service. Using these strategies, it is more likely that the confidentiality of clinical records will be maintained.

6.3. Become critical consumers of computer-based interpretive programs (Snyder, 2000)

Clinicians need to be aware whether the interpretation software is based on a valid test, if the authors of the interpretive software have established expertise, if reliability and validity studies of the software are available, if a user's guide is available, if the software is flexible enough to be revised and updated with new empirical information, if moderating effects need to be taken into account during interpretation, if information is provided as to the basis of the interpretive report (actuarial or theoretical?), and if additional information (e.g. statistical information, probability estimates that indicate the degree that a respondent's profile is consistent with prototypical patterns) is provided (Butcher, 1995, 2003; Moreland, 1985; Roid, 1985, as cited in Snyder, p. 56). If these kinds of information are not readily available, then clinicians should raise these concerns with the producer of the software.

Clinicians should base their test interpretation and impressions on their clinical acumen, as well as reliable and valid testing results, not solely on interpretations generated by computer software that are accepted at face value without critical examination. Simply because results come from a computer does not ensure the data's validity (Butcher, 2003). It is important for clinicians to interpret the test based on standard scores prior to reviewing the computerized interpretive report. In this fashion, clinicians may solidify their impressions based on their training, and they would be better able to judge the interpretive statements to determine the ones that accurately reflect a responder. Some clinicians faced with a complicated diagnostic picture may be tempted to rely on the computerized interpretive report too heavily, detracting from the overall quality of the assessment. Moreover, many of these interpretive reports come with diagnoses that may "fit" the pattern of response. It is not known how seeing these diagnoses early on during an assessment can influence the interpretive process.

For a thorough examination of common errors in using computerized assessment packages (e.g. not attending to norms, not attending to base rate information), and recommendations as to how to improve clinical decision-making (e.g. increasing one's awareness of research), readers are referred to Garb (1998). With regard to improving the education of clinicians in this area, Butcher et al. (2000), Butcher
(2003), and Snyder (2000) recommend that clinicians receive specialized training in interpreting computer test packages. It is concerning that, in a survey of APA-accredited clinical and counseling graduate training programs, Olson (2001) found that the graduate training in computerized testing was limited. It may be helpful for clinicians to initially be exposed to ethical concerns, potential judgment errors, and possible pitfalls in evaluating computer-generated reports through their assessment training during graduate school.

6.4. Know the psychometrics of the computerized measure

Clinicians must evaluate whether separate norms have been established for the computer version of an assessment measure they want to administer. If separate norms have not been developed, then several questions are raised. For example, what studies have been performed to determine equivalence? To what degree has equivalence been established? Clinicians who are unable to answer these questions risk violating ethical standards. A safe method is to initially treat a computer version of a paper-and-pencil test as if it were an entirely different measure, and not an alternate form of a questionnaire that has been empirically validated. In this vein, given that there are confounds to equivalence, clinicians should not simply extrapolate paper-and-pencil norms to assist in computer-based interpretation. Moreover, clinicians should not assume that research on a paper-and-pencil form of an instrument is applicable to an instrument’s computerized version. Nonetheless, clinicians often use research data from paper-and-pencil administered tests in such a fashion (Garb, 2000b), which is an unfortunate practice.

6.5. Monitor user qualifications

With regard to the possibility that unqualified persons may purchase or administer psychological tests, testing companies must be careful to scrutinize the credentials of the purchasers of their products. Furthermore, clinicians who utilize computerized psychological tests in their work must assume responsibility for their administration. This includes supervising staff who administer computerized tests to ensure they are adequately trained to respond to concerns that may arise during an assessment. Furthermore, though not explicitly stated in professional standards, competent practice should include not only appropriate credentials and familiarity with the psychometrics of the computerized tests or computerized interpretations, as mentioned above, but should also include a basic familiarity and experience with computers. We should expect no less of ourselves than what we expect of our clients regarding comfort and familiarity with computers.

7. Recommendations for research

As computerized assessment becomes more prevalent in clinical practice, there is a parallel need for continued research into its limitations and benefits. Examples of
possible areas for further research include improving the quality of assessment software, examining the accuracy of mechanical versus clinical models of decision-making, and continuing to examine effects of responder characteristics on computerized assessment outcome.

7.1. Improve the quality of assessment software

If clinicians and responders are to benefit from the use of computers in clinical practice, then improvements must continue to be made in the quality of the software, how this software is evaluated, and their degree of accuracy (e.g. Butcher et al., 2000; Garb, 2000a). Although it is beyond the scope of this article to thoroughly address this issue, Garb provides useful recommendations (specific to improving statistical prediction) for creating and appraising computer programs. In a more general sense, systematic research needs to be conducted to determine to what degree the interpretive software is reliable and valid, as well as whether the software is biased with certain populations or under certain conditions.

Although there have been surveys conducted in the area of ethics and computer-based assessment (e.g. McMinn, Buchanan, Ellens, & Ryan, 1999; McMinn, Ellens, & Soref, 1999), more needs to be done. There may be a distinct difference between what people know they should do and what they actually do in clinical practice. Also, more research needs to be done to adequately understand how viewing diagnoses from computerized interpretive reports can influence clinical decision-making. Likely, clinicians with extensive experience and/or those working with a textbook diagnostic picture are not adversely influenced by computerized interpretive report diagnoses. Less experienced clinicians and/or those facing a complex diagnostic picture may place more emphasis on diagnoses yielded by a computer-based test interpretation.

7.2. Examine the accuracy of mechanical versus clinical models of decision-making

Mechanical models of decision-making appear to be an extension of the statistical versus clinical prediction debate that was initially introduced by Meehl (1954). This debate continues today. In a recent meta-analytic review, Grove, Zald, Lebow, Snitz, and Nelson (2000) found that, on average, mechanical prediction was more accurate than clinical prediction (by about 10%). "Although research generally indicates that mechanical- or computer-based models for decision making outperform clinical decisions based on the same data, the superiority of mechanical methods is neither universal nor consistently large" (Snyder, 2000, p. 57). Snyder astutely observes that researchers should focus on ways to adaptively integrate both predictive approaches, rather than focusing on which is better (mechanical or clinical).

7.3. Continue to examine effects of responder characteristics on computerized assessment outcome

Many scales have been designed to assess computer-related constructs such as aversion and attitudes (see LaLomia & Sidowski, 1991, 1993, for an initial review,
although scales have been further researched and new ones developed since these papers were published). Once more research has been conducted to further define and expand their psychometric properties with additional populations of individuals, these scales may eventually enhance the efficacy of the testing process. Currently, discussion and observation may be most helpful with regard to evaluating a client's attitudes toward computers or their comfort level during a computerized assessment.

Although there is growing discussion in the literature with regard to how computer aversion and attitudes toward computers can influence a computerized psychological assessment, there is much that remains unknown. For instance, if it is indeed the case that a high degree of computer aversion can artificially inflate scores on a computerized measure of negative affect, then it is possible that those who enjoy working on computers may have artificially decreased scores on computerized measures of negative affect. Moreover, how would constructs such as computer aversion influence scores on measures of personality, measures of career interest, neuropsychological measures, and/or educational measures? Additionally, if constructs such as computer attitudes can influence achievement, as researchers such as Marcoulides (1988, 1991) indicated, are there certain kinds of achievement (mathematical, reading) that are influenced, and to what degree? Further research is needed to investigate these issues. Until these questions are answered empirically, clinicians should be aware that variables related to the human-computer interface may artificially influence results.

7.4. Examine the impact of culture on computerized testing

There are articles available on computer-related constructs in different cultures (e.g. Weil & Rosen, 1995), but more papers are needed to better understand how constructs such as computer aversion and attitudes towards computers may influence the administration and interpretation of computerized tests from a cross-cultural standpoint. It is possible that response patterns from populations relatively unfamiliar with traditional "bubble" forms (e.g. immigrants, the elderly) may be more reliable and valid if gathered through a computer format.

8. Concluding remarks

Computerized psychological testing is a promising area for the clinician and the researcher, but it is also one that continues to present a number of ethical challenges. The focus of this paper was to increase clinicians' awareness of ethical issues that they might face as they incorporate technology into their practice. It is hoped that the recommendations outlined above will encourage clinicians to seriously consider the strengths and limitations of computerized assessment on the quality of their clinical work and to encourage additional research in this important practice domain.

The authors would like to thank an anonymous reviewer for this observation.
Acknowledgements

The authors would like to thank Jessica Kaster, M.S. for assistance with helpful editorial comments made on a previous draft of this manuscript.

References


Garb, H. N. (2000a). Computers will become increasingly important for psychological assessment: not that there’s anything wrong with that! Psychological Assessment, 12, 31-39.


