PERGAMON

Computers in Human Behavior 17 (2001) 295–314

Computers in Human Behavior

www.elsevier.com/locate/comphumbeh

From Eliza to Internet: a brief history of computerized assessment

J. Epstein*, W.D. Klinkenberg

University of Missouri School of Medicine, Missouri Institute of Mental Health, 5400 Arsenal Street, Saint Louis, MO 63139, USA

Abstract

Three decades of research into the equivalency of traditional and automated assessments have culminated in the growing popularity of Internet-based surveys. This article briefly reviews the history of computerized assessment and compares and contrasts it to the current research on Internet-based surveying. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Assessment; History; Internet; Equivalency

1. Introduction

From nearly the moment computers were made available to academicians, researchers have been investigating the effectiveness of automating psychological assessments. From simple true/false questionnaires to complex psychological histories, countless measures have been computerized. Not only has this computerization aided clinicians, it has also provided researchers a vast new territory to explore. Research has been conducted regarding the reliability, validity, user acceptance, and cost-effectiveness of these automated assessments. Three decades of research examining the equivalency of computerized versions of traditional interviews and psychological tests has led to the conclusion that automation is accompanied by many more benefits than problems.

This article begins with a brief review of computerized interviewing and test administration along with a discussion of their associated strengths and weaknesses. Following this review is an overview of how researchers are using these previous

* Corresponding author. Tel.: +1-314-644-8814; fax: +1-314-644-7843.
E-mail address: epsteinj@mimh.edu (J. Epstein).

PII: S0747-5632(01)00004-8

studies as a platform from which to conduct new research into the viability of using the Internet to collect psychological data.

2. Computer assisted interviewing

One of the first attempts to simulate the behavior of a psychotherapist was in 1966 with a program called Eliza (Weizenbaum, 1976). Eliza acted as a Rogerian therapist, asking the user to explain his/her feelings. Although the goal of the program was to "demonstrate that the communication between man and machine was superficial" (Nadelson, 1987), Weizenbaum (1976) was surprised to find that people enjoyed using Eliza and actually attributed human-like feelings to the program.

Eliza has been hailed as one of the first programs able to past the "Turing Test". A pioneer of Artificial Intelligence, Turing (1950) developed a simple model for determining if a computer program was intelligent: could the program fool its users into believing they were interacting with a human instead of a machine? Participants would sit at a remote terminal and receive and send messages to either a computer program or human confederate. After a period of interaction, the participant was required to judge whether the responses were generated by a computer or a human.

Like Eliza, Parry (Colby, 1975; 1981) is another psychiatrist-authored computer program which has passed the "Turing Test." However, unlike Eliza's supportive and friendly tone of a caring psychotherapist, Parry was programmed to be hostile and defensive, mimicking the behavior of a paranoid individual.

Though both Eliza and Parry were experiments in Artificial Intelligence designed to imitate human behavior, the results of research with these programs demonstrated that human/machine interactions were not only possible, but deserved further attention. One avenue researchers began to explore was the development of computer assisted interviewing (CAI) software. Since the early 1980s, scores of such programs have been developed including ones designed for assessments of depression (Carr, Ancill, Ghosh, & Margo, 1981), phobias (Carr & Ghosh, 1983), fundamental problem areas of psychological functioning (McCullough, 1983), substance abuse (Davis, Hoffmann, Morse, & Luehr, 1992), and suicidal risk (Greist, Gustafson, Strauss, Rowse, Laughren, & Chiles, 1973). Perhaps one of the most ambitious CAI programs is the computerized interview developed by Angle (1981). It takes approximately four to 8 h to complete, asks nearly 3500 questions, and provides the clinician with a full report of a client's problem areas.

2.1. Criticism of CAI

Many clinicians have argued that CAI software cannot possibly conceptualize a person in his/her totality. These clinicians feel that the computer cannot tell when the client is having difficulty expressing his/herself, cannot discriminate between error and pathological responses, and cannot record the "rich variety of behavioral data" (e.g. posture, eye-contact, etc.) during an interview (Space, 1981). Because of these limitations, many clinicians have considered CAI software to be cold and

inhumane (Erdman, Klein, & Greist, 1985). Additionally, because using CAI software requires some degree of literacy and computer skills, persons with limited educational experience may not be able to adequately use these programs.

Clinicians have also pointed to their own strengths when criticizing CAI software. These clinicians feel that they are able to be much more flexible than the programs can be. For example, clinicians can alter their tone of voice or sophistication of language, while CAI software is limited by how it has been programmed (Erdman et al., 1985).

2.2. Advantages of CAI

Support for CAI programs has come mainly from researchers who point to the limitations of human interviewing skills. Clinicians have been shown to be very inconsistent in the way they administer interviews, often forgetting to ask important questions (Angle, 1981; Erdman et al., 1985). In contrast, proponents of CAI software point to the fact that computerized interviewing programs are not affected by factors such as the mood of the interviewer or the time of day (Space, 1981). In fact, as Erdman et al. (1985) have stated, the computer "does not get tired, angry or bored. It is always willing to listen and give evidence to having heard. It can work at any time..." Other advantages of CAI programs are that they allow the client to proceed at his/her own pace, they allow for convenient storage and access of past interview sessions, and they are more cost-effective than traditional interviews (Space, 1981).

Researchers have attempted to mediate the controversy surrounding CAI programs by empirically evaluating the effectiveness of the software. Mathisen, Evens, and Meyer (1987) developed a computerized version of the Diagnostic Interview Schedule which provides the clinician with a three to four page diagnostic report and potential diagnoses. Mathisen et al. (1987) used the program with 135 psychiatric patients and asked 23 psychiatrists to evaluate the reports. Psychiatrists rated 61% of the reports as being somewhat or very accurate while only 11% of the reports were rated as not at all accurate. Additionally, 57% of the patients indicated that they liked the interview experience while only 2% stated that they did not like it. Seventy percent of the patients said that the program was very easy to use and only 4% said that they would be more embarrassed by using the computer than by having a traditional interview. Other researchers (Davis et al., 1992; Millstein, 1987; Greist et al., 1987; Quintar, Crowell, & Pryor, 1982; Slack & Slack, 1977) have found similar results.

2.3. Questions remaining for computerized assisted interviewing software

2.3.1. Population differences in response to CAI

Further research is needed to examine how different populations respond to CAI software. Little research has examined the interaction between age, gender, or culture and automated interview procedures. Such investigations may lead to more sensitive measures.

2.3.2. AI based CAI programming

Although Eliza and Parry provided the impetus for the development of CAI software, there has been a general disconnect between the models these early programs used and current interviewing software. Most of the available programs are either ports of existing paper-and-pencil interviews or static models developed solely for computer administration. Collaboration between psychologists and experts in Artificial Intelligence may lead to the development of dynamic interviewing programs which can reliably and accurately assess a wide array of individuals across a broad continuum of disorders. Finally, developments in speech recognition may eventually make it possible for CAI software to allow interviewees to communicate via natural language.

3. Computer assisted testing

In addition to computerized interviewing, researchers have also examined whether computerized and traditional test administration are equivalent. The simplest types of computer assisted testing (CAT) programs have the computer display the questions of a test and ask the client to enter his/her response (Rozensky, Honor, Rasinski, Tovian, & Herz, 1986). These programs attempt to make the computerized test-taking experience as identical to the traditional test-taking experience as possible.

Other types of CAT programs use an "adaptive testing" procedure. Roid and Gorsuch (1984) believe that this adaptive testing approach allow computers to "go beyond a simple page-turning function." In this procedure, the computer presents an item, scores the response, and then selects the next appropriate question (Green, Bock, Humphreys, Linn, & Reckase, 1984). This method is similar to that of many non-computerized intelligence tests.

Conventional tests have either a large set of items with one central level of difficulty or a limited range of items covering a wide range of difficulty levels, but adaptive tests have a large number of items which cover a complete range of levels of difficulty (Weiss, 1985). By comparing the test-taker's performance to the ratings of the items, the computer is able to "individualize" any testing session by dynamically presenting only as many items needed to insure that an exact measurement of the trait being studied is obtained (Burke, 1993; Weiss, 1985). Accordingly, "the final trait estimate is independent of the particular set of items given to the examinee (and) scores can be compared even though the examinees might have been administered completely different sets of items" (Burke, 1993).

There are several types of adaptive tests. In the "two-stage" adaptive test, an individual takes a pre-test called a routing exam. Based on the performance on this test, the individual is then administered a test which approximates his/her level of mastery of the subject. The "pyramidal" adaptive test involves "layers" of test items. An individual begins at the top level with an item of average difficulty. Under each item are two more sets of items; one that is a bit easier and one that is slightly more difficult. Depending on the individual's response, the computer will branch to

either the easier or more difficult set of questions. The "flexilevel" adaptive test is largely similar to the pyramidal approach except that the flexilevel has only one question per level (instead of an entire set per level). The "stradaptive" approach has a larger item pool and administers the "most discriminating unadministered item" in the pool to an individual's response (Weiss, 1985). Finally, the "countdown" approach to adaptive testing involves administering items until it is clear that the individual either will or will not reach the criterion cutoff point for the given scale (Butcher, Keller, & Bacon, 1985).

Since the 1980s, advances in computing power and new theories of test development have led to entirely new methods of computerized testing. For example, (Bennett et al., 1999) describes a test type called Generating Examples (GE) in which items may have multiple solutions and respondents are required to provide one or more of these solutions. Automated scoring is possible by applying an underlying construct to the user-provided solution. A similar type of CAT program is described by (Bennett, 1999). This program allows instructors to input general algorithms from which to construct test items. Based on these algorithms, the program will automatically generate a nearly infinite number of questions which will allow the assessment of the underlying cognitive principle. More dramatic advances in CAT programming come in the form of tests designed to grade essay responses (Bennett, 1999) and other types of open-ended responses (Bennett, Steffen, Singley, Morley, & Jacquemin, 1997). Finally, computerized simulations are now being used to assess students' ability to utilize their knowledge. These simulations are more performance-based types of exams which purportedly mirror real-life scenarios in which learned knowledge must be applied (Bennett, 1999).

Many of the advancements in CAT programming has come from the Educational Testing Service (ETS, 2000). Founded in 1947 to assume the testing activities of the American Council on Education, this group rapidly became the world's largest testing organization In the 1998–1999 school year ETS was responsible for administering and scoring nearly 2.5 million SAT tests in the United States and more than 11 million tests worldwide. A large proportion of these tests are computerized. Indeed, much of the research on automated testing has originated in ETS laboratories (Bennett, 1999).

3.1. Criticism of CAT

Critics of CAT technology believe that computerized test administration is impersonal and feel that it introduces too many irrelevant factors into the testing situation. Factors such as lack of familiarity and anxiety with computers are often cited (Hofer & Green, 1985).

Perhaps the greatest criticism against CAT programs, however, centers on the issue of equivalency. Honaker (1988) stated that many previous researchers have failed to consider whether a computerized test produces the same rank ordering and distribution differences as its traditional counterpart. Burke (1993) concurs, stating that in analyzing CAT programs, it is not enough to say that they produce similar mean scores as paper-and-pencil versions, but it is equally, if not more important to

determine if the two versions produce similar frequency distributions. That is, does a single individual rank similarly across the two methods of test administration?

Additionally, Honaker (1988) argued that taking a test via a computer terminal is such a dramatically different experience for the respondent that it guarantees that the two types of administration will not be equivalent. Similarly, Bennett (1999) has stated that in addition to the actual constructs being measured, automated assessments introduce additional factors which need to be carefully evaluated. Issues such as, the interactions between the test content, the test design, and the user interface are all important considerations. For example, in mathematics, solutions are often derived using complex scribblings, diagrams, etc. If a computerized assessment program restricts this type of problem solving by forcing examinees to progress through solutions in a strict linear manner, the very nature of the assessment changes.

3.2. Advantages of CAT

Studies which have compared results between traditional and computerized tests, however, have found few psychometric differences between the two types of administration. For example, Elwood and Griffin (1972) examined the effects of computerization on IQ scores. They obtained a correlation of 0.95 between scores on the computerized and the traditionally administered versions of the WAIS, indicating that participants did, in fact, maintain the same rank ordering across the two methods. In fact, whether it be performance based tests such as the assessment of children's reading skills (Evans, Tannehill, & Martin, 1995), objective personality tests such as the MMPI-II (Pinsoneault, 1996), projective personality tests (Rasulis, Schuldberg, & Murtagh, 1996), or neuropsychological assessments such as the category test (Choca & Morris, 1992), researchers have found remarkable psychometric equivalence between traditional and computerized test administration (Campbell et al., 1999).

Although much has been written describing anxiety surrounding the perception that computers are responsible for our rapidly changing and sometimes dehumanizing culture (Wilson, 1999), evidence demonstrates that examinees typically do not mind computerized administration of testing materials. In fact, "clinical populations do not routinely suffer from cyberphobia" (Pinkerton & Raffoul, 1984). Research has demonstrated that most respondents report enjoying computerized tests more than the traditional versions of the same test (French & Beaumont, 1987; Hile & Adkins, 1997; Rozensky, et al., 1986). Other research indicates that rather than feeling anxious while interacting with computer programs, many respondents tend to provide more honest answers than individuals who received face-to-face interviews (Barak, 1999; Burke, 1993; Martin & Nagao, 1989).

Other advantages of Computer Assisted Testing include: (1) disposable materials such as paper are saved, reducing both storage and environmental strain; (2) standardization of timing and instructions are assured; (3) scores can be rapidly computed and immediate feedback can be given to examinees; (4) test security is improved; and (5) scores can be automatically added to databases for dynamic adjustment of norms (Barak, 1999).

3.3. Questions remaining regarding computerized assessment

3.3.1. Interaction between test content and test design

Research has generally found high correlations between computerized and paper-and-pencil test administration, although these correlations have rarely approached unity. Bennett et al. (1999) argue that because the correlation between a Generating Examples (GE) test and the GRE General test was 0.82, the two are not identical and therefore tap differing cognitive processes. It is true that high correlations between scores obtained from the two methods do not necessarily indicate similar construct validity; mean scores could be very different between the two methods, even though participants maintain the same rank order. More methodologically rich studies that use a "gold standard" (a third measure with good validity) or multi-trait multi-method matrices would be other useful steps in evaluating the equivalency of the two methods. Even if it is true that different underlying constructs may be measured by computerized versus traditional testing methods, researchers must carefully examine potential interactions between a test's content and both its method of eliciting responses and its user interface. Furthermore, additional research should be conducted on how pre-assessment tutorials affect final performance (Bennett & Bejar, 1998).

3.3.2. Computer anxiety

It may actually be that professionals administering computerized assessments are more fearful of technology than their clients. In one study of licensed psychologists, nearly half rated themselves as either mildly or highly technophobic (Rosen & Weil, 1996). In a meta-analysis of compuer-phobia, Chua, Chen, and Wong (1999) found that there are few correlates of either gender or age based computer fear and that such anxieties are generally state, not trait based. This assertion is confirmed by Wilson (1999) who demonstrated that with brief counseling sessions, computer-phobic individuals rapidly became willing to use new technology. However, on a cognitive task of pattern recognition, older adults demonstrated significantly higher levels of computer anxiety than did younger participants (Laguna & Babcock, 1997). Therefore, it seems prudent to assess computer anxiety in relation to the tasks being administered. It may be that some older people recognize they have difficulty performing on cognitive measures (because of age-related declines) and resultant differences in anxiety may actually be due to an interaction between task and age. Alternately, the relationship between performance and anxiety may be completely mediated by the amount of computer experience participants have. Further research is needed to both examine these types of relationships and ways to reduce anxiety professionals feel regarding the adoption of new technology.

3.3.3. Consequences of computerized scoring

Teaching practices may be influenced by the development and use of new computerized testing and scoring procedures. For example, although research has demonstrated that both humans and computer programs typically grade essays similarly (Burstein et al., 1998), the way in which the essays are scored may

actually influence classroom instruction. Because some essay assessment software examines both deductive reasoning and grammatical structure and others completely ignore mechanics (Bennett, 1999), research should be conducted to examine the influence (intended or not) such programs may have on instructional methods.

3.3.4. Generalizabilty of simulations

Further research is also needed to determine if simulations accurately probe the reasoning and other skills needed to perform in real-world situations. As computing power continues to increase, the realism of simulations are bound to increase. However, because of the tremendous expense of developing such assessments, research is needed to determine if proficiency in "virtual worlds" consistently translates to proficiency in "real world" settings (Bennett, 1999). Again, examinations of learner characteristics that may interact with success in simulations may be another fruitful area of study.

4. Psychological research on the Internet

After three decades of investigation, both researchers and clinicians generally accept computerized test administration as being a valid and reliable alternative to traditional methods. In fact, because of its generally high reliability, decreased costs, and greater client acceptance, computerized test administration may often be preferred. While research in this area will continue to be published, many researchers are now focussing their efforts on evaluating the equivalency between traditional and Internet-based assessments.

Although web-based data collection has been described in the literature as far back as 1996 (Welch & Krantz, 1996), published reports of psychological data collected via the Internet are still relatively rare (Pettit, 1999). In fact, only a handful of studies have been published which report the equivalence of web versus paper-and-pencil administration (Stanton, 1998).

Despite the relative dearth in published studies, there are countless surveys posted on the Internet. Many of these studies, however, are not of an academic nature. For example, there are customer satisfaction surveys, political polls, and relationship assessment measures. In fact, some of the most widely-used Internet-based surveys are ones in which individuals can perform "self-assessments". Two popular Internet sites, The Spark (www.thespark.com) and Emode (www.emode.com), offer a variety of surveys which provide users with feedback on everything from their intelligence to their personality and mental health.

Many clinicians may strongly object to the idea of personality tests being offered for free via the Internet. However, there are dozens of other professionals who are using the web as a way to collect legitimate psychological data. The American Psychological Society maintains a web-page (psych.hanover.edu/APS/exponnet.html) which provides a listing of on-going Internet data collection efforts. The list includes such domains as clinical psychology (e.g. trauma and social support), cognition

(e.g. decision making and memory), developmental issues (e.g. child discipline practices), and social psychology (e.g. partner selection and friendship). To date however, few of these studies have been published in peer-reviewed journals.

4.1. Problems associated with web-based data collection

Although researchers have identified many potential problems which may be associated with on-line data collection, the difficulties may be classified into three major categories: (1) sampling, (2) lack of control over the experimental environment, and (3) data integrity.

4.2. Problems associated with sampling

Obtaining a representative sample via Internet surveys may be quite difficult. Unlike telephone directories, the web has no central registry of users (Kaye & Johnson, 1999), making it very difficult to establish a sampling frame. Additionally, the average Internet-user may be quite atypical from the sought-after population (Davis, 1999; Richard, 2000; Stanton, 1998). For example, individuals from a low socioeconomic status are generally underrepresented in on-line samples (Michalak, 1998). In fact, the average Internet user is a married white male 26–30 years old with a college degree earning between $50–74,000 annually (Georgia Institute of Technology, 2000). Overall, Krantz and Dalal (2000) state that the external validity, or generalizability of Internet studies is typically rather low because researchers do not, or cannot, make efforts to ensure that samples used are randomly selected and therefore representative of the larger population being studied.

Another problem associated with sampling includes the translation of concepts into other cultures and languages. For example, how can a researcher possibly expect a Zimbabwe national to be able to report annual income in terms of American dollars (Richard, 2000)? Similarly, some types of content may be treated very differently across different cultures and make interpretations of results difficult (Reips, 2000).

Other sampling issues include self selection bias (Smith & Leigh, 1997) and participant motivation (Stanton, 1998). Because many researchers have traditionally used undergraduate psychology students as participants, and these students typically obtain course credit for their participation, they may feel some pressure to see a study through its end. With web-based surveys, however, there are no such social/academic pressures (Reips, 2000). Finally, it may be difficult to establish response rates because it is nearly impossible to tell how many people have seen announcements for the study but decline to participate (Kaye & Johnson, 1999).

4.2.1. Problems associated with control over the experimental environment

Researchers lose almost all hope at controlling the experimental environment when they decide to collect data via the Internet (Davis, 1999; Smith & Leigh, 1997; Stanton, 1998). Not only is there little assurance as to who will participate, there is absolutely no way of knowing the psychological status of the participants. Like mail

surveys and some self-completed questionnaires, people may be distracted, fatigued, rushed, or even intoxicated while completing the survey.

The type of computer and version of Internet browser software will also vary from individual to individual (Pettit, 1999; Schmidt, 1997). Not only will some computers not be able to adequately display certain types of graphic and multimedia content, not every Internet browser software supports all types of programming. Therefore, not only may a survey appear differently across different machines, it may also function differently across different software packages. It is virtually impossible to test an Internet application with every available platform and software package and thus, nearly impossible to assure that all users will have the same test-taking experience.

4.2.2. Problems associated with data integrity

Once users have submitted data to an Internet survey, researchers have additional difficulties. First, because the data are sent (and often stored) on-line, there is a potential for breaches in security (Pasveer & Ellard, 1998; Schmidt, 1997). Computer hacking can make the survey and its results available to unintended persons (Smith & Leigh, 1997). Although such breaches in security are possible, the likelihood is remote and fairly easy to prevent by employing methods such as using secure web sites, encrypting data before it is transmitted, and storing data in secure directories behind firewalls (Barak & English, 2001; Schmidt, 2000). Overall, Barak and English argue that "there is no evidence that test data are more secure if paper-and-pencil tests are used and materials are archived in standard office cabinet files or drawers than on an Internet server...there appears to be no greater risk for data collected on the Internet to be illegally obtained than any other electronic data or data saved on traditional paper materials".

While hacking data is fairly easy to prevent, researchers have little control over other aspects of their data. For example, it is nearly impossible to prevent hoax, or disingenuous responses to the survey (Davis, 1999; Michalak, 1998). An equally problematic issue is that of multiple submissions from a single individual.

4.3. Benefits of web-based data collection

Oddly enough, the benefits of web-based data collection closely mirror its associated problems. That is, Internet surveys offer advantages in terms of: (1) sampling, (2) control over the experimental environment, and (3) facilitation of data collection and analysis.

4.3.1. Benefits associated with sampling

The number of individuals using the Internet is growing at an exponential rate. In April of 1999 there were approximately 92 million Internet users in the United States and Canada alone, representing an increase of about 13 million users in a period of just 9 months (CommerceNet, 2000). But this growth is not limited to North America. In fact, even developing nations are showing similar patterns of expansion (Schmidt, 1997). With such a huge population of users, the Internet allows for access

to large samples of individuals which had, until now, been inaccessible (Michalak, 1998; Pasveer & Ellard, 1998), thus aiding cross-cultural research (Smith & Leigh, 1997). However, because there are so many people using the Internet, researchers are also able to target very specific populations such as youth, gays and lesbians, or trauma survivors (Michalak, 1998; Schmidt, 1997). Researchers also have the ability to limit who will be granted access to their survey by distributing Personal Identification Numbers or having their server deny access to certain domains (e.g. restricting access to allow use only by students from the researcher's home university; Schmidt, 2000). To control for self-selection, researchers may recruit participants via multiple means, providing a separate URL for each method. By examining differential access to these URLs, researchers can identify if some type of systematic self-selection processes influenced participation in the study (Reips, 2000).

4.3.2. Benefits associated with control over the experimental environment

As with other computer assisted testing programs, individuals may find Internet surveys less threatening and more interesting than traditional paper-and-pencil tests (Stanton, 1998). The very nature of Internet surveys offers increased anonymity and decreased demand characteristics (Davis, 1999; Pasveer & Ellard, 1998; Smith & Leigh, 1997). Because Internet surveys may be perceived as less threatening and more anonymous, they may facilitate greater honesty and self-disclosure. Also, because the experimenter is not present with the participant, observer biases may be eliminated (Davis, 1999). Finally, Internet surveys may be designed to be dynamic — changing with users' responses. This type of interactivity is likely to increase participant interest and motivation (Schmidt, 1997).

4.3.3. Benefits associated with data collection and analysis

Internet surveys allow participants to enter data at their own, not the experimenter's, convenience (Davis, 1999). Consequently, large numbers of participants may be collected in a short amount of time. For example, Birnbaum (2000) was able to collect 1224 responses in a period of 4 months, nearly tripling the rate at which he had been collecting subjects via traditional means.

Web-based surveys can be designed to check responses before they are entered, thus assuring that the data are well structured and free from missing values or out-of-range responses (Stanton, 1998). Furthermore, data entry errors are eliminated because responsdents' answers may be entered directly into an analyzable database, completely eliminating the need for a separate process of data entry (Michalak, 1998; Pasveer & Ellard, 1998).

As noted earlier, critics of Internet surveys often express concerns about data integrity issues. Even though no single method of assessing data integrity is likely to emerge, several, imperfect, options do currently exist to evaluate the quality of responses. As with traditional surveys, it is possible to examine pairs of items for inconsistent responses. For example, it would be wise to be wary of the validity of the responses from men who say they are pregnant or from respondents who deny drug use in one section, then describe daily marijuana use in another. Other options that take advantage of the technology include using cookies (small text files written

to a participant's computer and later read by the remote server) to monitor the characteristics of respondents. Cookies could be used to track the number of times any given computer has been used to access the survey, could be encoded with key demographic responses from earlier survey completions, or could be encoded with an ID number (Musch & Reips, 2000; Schmidt, 2000). Another option that has been used to discourage multiple completions from a single user is requiring a potential respondent to identify a valid email address before beginning the survey. Finally, once suspicious data are identified, decisions can be made about which cases and/or values should be excluded from the statistical analyses.

5. Previous Internet-based research

Combined, the benefits of Internet-based surveys allows researchers to save time and money (Davis, 1999; Nicholson, White, & Duncan, 1998; Schmidt, 1997). Consequently, although there are currently few published examples, there are likely to be increasing numbers of psychologically oriented research studies conducted on the web. Below, is a review of several previously published studies. These examples are organized into five separate categories: (1) e-mail surveys, (2) descriptive surveys, (3) validations based on previously published findings, (4) validations based on a traditional comparison group, and (5) theoretical validations.

5.1. E-mail surveys

Researchers have extended the methods of traditional survey administration to the Internet by conducting e-mail surveys (Anderson & Gansneder, 1995). Using this methodology, researchers send e-mails to either random or targeted e-mail lists and ask readers to respond to the survey via e-mail. Although this can be a quick and easy method of sampling, it is at the cost of highly unformatted and unpredictable responses. Researchers using e-mail surveying techniques must spend a considerable amount of time cleaning their data before data analysis can begin.

Mehta and Sivadas (1995) found that traditional mail-based surveys and e-mail surveys tend to have similar response rates. Swoboda, Muhlberger, Weitkunat, and Schneeweiss (1997) successfully used an e-mail survey to assess Internet users' perceptions of "Future Risks". The researchers set up a program to automatically capture the e-mail addresses of individuals who posted messages to 200 randomly selected Internet newsgroups. In 2 weeks, 8859 addresses were collected and 1713 individuals (20.8%) responded to their e-mail query.

5.2. Descriptive surveys

Instead of using e-mails, other researchers have attempted to collect data by posting surveys on the Internet. Using standard HTML programming, researchers can quickly create form-based questionnaires. Depending on their programming sophistication, researchers may either have the responses to the survey returned via

e-mail or, more elegantly, written directly to an analyzable database. The simplest types of Internet studies merely attempt to provide descriptive information based on the results of respondents' responses. Musch and Reips (2000) argue that data obtained from such Internet surveys may be of higher quality than data collected via traditional means because Internet participants are likely to be less prone to the effects of social desirability and typically give more elaborate and complex responses. Furthermore, descriptive surveys collected via the Internet may provide instructive results because they are likely to contain responses from individuals not typically included in traditional surveying procedures.

For example, Kaye and Johnson (1999) created a 20-item Likert-type questionnaire to assess Internet users' political preferences. After a month of data collection, the researchers received, via e-mail, 306 completed surveys (average age = 31.2, 75.5% male, 60% college education or higher). The researchers posted announcements about their survey in various forums and asked respondents to indicate where they had first heard about the survey. Results indicated that 32% responded from announcements in chat forums, 23% responded from links posted on other web sites, and 17% responded from both postings in newsgroups and on listserves.

Nicholson, White, and Duncan (1998) also conducted a descriptive Internet survey. Their goal was to determine if there is a large population of successful, well-functioning adults who are casual users of illicit drugs. They created a four-part survey (demographics, drug experiences, legal history, and general well-being) and posted it on the Internet for 2 weeks of data collection. The researchers posted announcements of their survey on Usenet newsgroups (i.e. all of alt.drugs.* and rec.drugs.*). A total of 276 responses were received (average age = 32.34, 78% male, 93% white). Although this study did use some sophisticated programming (e.g. only assessing respondents' behaviors with drugs for which they had indicated having used), the researchers allowed respondents to skip individual items, thereby leaving the researchers with missing data in several places.

5.3. *Validations based on previously published findings*

Instead of attempting to merely describe a population and their behaviors, other researchers have attempted to collect data via the Internet and demonstrate its validity by comparing results to those found in the literature.

Michalak (1998) conducted a web-based study to assess Seasonal Affective Disorder (SAD). In 4 months of data collection in which an on-line survey was publicized on various newsgroups and search engines, 425 surveys were completed. Results indicated that 30% of the sample could be classified as having SAD. Although Michalak found a low correlation between latitude and seasonal depression, several results did confirm previous research findings including an inverse correlation between SAD and age and a high prevalence of sub-clinical manifestations of SAD.

Stones and Perry (1997) conducted a similar survey to assess the incidence of panic attacks among Internet users. Publicizing their survey in newsgroups and on search engines, their 5-week data collection period yielded a total of 330 respondents

(64% female, 51% from the 21–39 age group). The researchers reported a 60% hit to completion ratio. Responses to the survey were congruent with symptomatology reported in the literature.

Richard (2000) conducted sophisticated analyses of responses to his web-based Post-Traumatic Stress Disorder (PTSD) survey. Publicizing his survey on search engines and a link on the International Society for Traumatic Stress Studies home page, his 10-month data collection period yielded a final sample of 260 completed surveys (72% female, 89.6% Caucasian, 61% with college degree or higher, from 19 different countries). Using confirmatory factor analyses, Richard demonstrated that despite his heterogeneous sample, participants' patterns of responding to the survey matched the same factors identified in the literature.

5.4. Validations based on a traditional comparison group

Instead of comparing results based on previous literature findings, several other researchers have conducted their investigations using both an Internet and a laboratory sample. This type of validation study is quite analogous to studies researchers have performed when trying to validate CAI and CAT software. Generally, attempts at establishing the convergent validity of web-based studies have been quite successful, with data collected in the laboratory showing the same psychometric properties as data collected via the Internet (Krantz & Dalal, 2000).

Buchanan and Smith (1999a) posted a self-monitoring questionnaire on the web and recruited participants via postings to newsgroups and solicitation of links on other web sites. Over a 2-month data collection period, the researchers obtained a total of 963 completed surveys (51% male, average age = 32). The researchers also collected data from a sample of 224 undergraduate volunteers (16% male, average age = 27). Despite the relative differences in demographics between the two samples, confirmatory factor analyses demonstrated that both the web-based sample and the undergraduate sample had identical factor structures.

Pasveer and Ellard (1998) obtained similar results in their study of self-trust. A sample of 429 Internet users were recruited via personal e-mails, links from other web pages, and listings on search engines. Researchers also obtained an additional sample of 760 volunteer undergraduate psychology students. Both groups took a 20-item 5-point Likert-type questionnaire on self-trust. Results were remarkably similar to those found by Buchanan and Smith (1999a): despite marked differences in demographics, the two samples' responses to the questionnaire were virtually identical in terms of factor structure.

Other researchers have conducted similar studies, using different measures. For example, Davis' (1999) investigation on rumination, Smith and Leigh's (1997) study on human sexuality and sexual fantasies, Bailey, Foote, and Throckmorton's (2000) investigation of sexual behaviors and attitudes, Stanton's (1998) survey on fairness in the workplace, Birnbaum's (2000) decision-making experiments, and Krantz, Ballard, and Scher's (1997) study on the determinants of female attractiveness all demonstrated that data collected via the web is virtually indistinguishable from data collected in the laboratory.

5.5. Theoretical validations

Although Buchanan and Smith (1999a) demonstrated congruence between their web-based measure of self-monitoring and laboratory sample, in an additional study (Buchanan & Smith, 1999b) they sought to demonstrate that their web survey actually measured the construct it purported to measure. Based on the literature's formulation of the concept of self-monitoring, the researchers identified groups of individuals who should theoretically be high or low in self-monitoring. They reasoned that actors and obese people would be high self-monitors and shy people and individuals interested in environmental causes would be low self-monitors. The researchers then identified newsgroups in which each of the identified groups might participate and sent announcements to these groups asking for volunteers to take their web-based survey of self-monitoring. Different URLs were used to track the newsgroups from which participants came. Results of the study supported the researchers' hypotheses. Participants from the high self-monitoring newsgroups did score significantly higher than those from the low self-monitoring newsgroups.

5.6. Research on Internet research

Krantz and Dalal (2000) reviewed a large portion of published research conducted via the Internet and Musch and Reips (2000) conducted a survey of researchers who have conducted Internet studies. Taken collectively, the findings of these two separate articles give an interesting overview of Internet research.

Among the findings include generalizations about gender, ethnicity, country of origin, and age of research participants. Typically, most respondents to web-based studies have been white North Americans with an average age of 35.2 (S.D. = 9.5). Europeans have been consistently represented at about 10% of all participants. No clear-cut gender trends are available, with percentages varying widely from study to study. From the practical standpoint, web-based studies have required an average of 22 min to complete with a range from between five and 90 min. The average total number of participants is 427 with a range from between 13 and 2649. On average, 68% of all visitors to a given web survey completed it, with a range from seven to 100%. Finally, researchers reported knowing of no attempts to breach the security of their survey or their data.

6. Conclusions and future directions

The preceding sampling of Internet-based research surely represents just the beginning of a new phase of investigations into the benefits of computerized assessments. Just as researchers have spent the past three decades investigating reliability and validity issues of CAT and CAI software, we are likely to see many years of research which examines the benefits and problems associated with Internet-based assessment.

Despite the fact that research conducted via the Internet is computer-based, many of the previous issues surrounding CAT and CAI software are no longer relevant. For example, the issue of end-users suffering from "cyberphobia" is a moot point when it comes to Internet-based studies. Not only has previous research concluded that most users actually prefer taking assessments via the computer, but Internet users represent a population who voluntarily seek to use their computers. Similarly, because previous research has demonstrated that translating a measure into a computerized format does not necessarily change its reliability and validity, new Internet-based investigations need not be so focused on demonstrating the equivalency of paper-and-pencil instruments to computerized versions that are identical in every other way. However, there is still much to be learnt about the equivalency of instruments when the testing method is altered with computerization (e.g. using adaptive testing methods). Further, there is a need for more research exploring the impact that greater respondent anonymity has on disclosure with computerized versions of traditional measures.

Perhaps the biggest challenges for researchers in designing Internet-based studies will be examining ways to deal with problems associated with data integrity and sampling. Clearly, Internet users represent a unique population of individuals who may or may not be representative of the general population on a number of important psychological variables. Similarly, because individuals from around the world have access to any content on the Internet, researchers must learn how to adequately handle respondents from outside the intended population.

Additionally, researchers using the Internet have faced the new obstacle of having limited ability to help respondents who have adverse reactions to participating in their studies. Such effects could be anything from mild distress felt after self-disclosing personal information to the more severe case of re-awakening suicidal ideations. New standards for ethics and guidelines for research may be needed to help establish practices for how to effectively debrief Internet research participants.

As computing power continues to escalate, and the Internet evolves even further than it already has, researchers should begin to examine how changes in society as a whole, and the Internet in general, might affect the outcomes of Internet-based studies. For example, one untoward effect of an increasing number of Internet-based studies might be difficulty recruiting participants as competition from other Internet studies increases. Similarly, researchers should examine methods of recruiting individuals into research studies and potential interactions between recruitment and content. For example, Buchanan (2000) notes that recruiting via newsgroup announcements may contaminate results because of the strong possibility that participants will discuss the study on the newsgroup. Meta-investigations of how the Internet is being used and how it is changing will provide a unique opportunity for collaboration between multiple professions including computer scientists, social psychologists, and clinicians.

Similarly, web-based research has opened the door to a new era of cross-cultural investigations. Sadly, few researchers have taken advantage of this new promise. A notable exception is work by Pagani and Lombardi (2000) which examined cultural and geographical difference in the perception of facial determinants of surprise.

Perhaps it is fitting that one of the first cross-cultural Internet studies examined surprise. There are, no doubt, many surprises to come in the field of Internet-based research. Ideally, however, with its global and always-available presence, Internet-based research will actually open the doors for researchers to gain a better understanding of humanity at large.

References

Anderson, S., & Gansneder, B. (1995). Using electronic mail surveys and computer monitored data for studying computer-mediated communication systems. *Social Science Computer Review*, *13*(1), 33–45.

Angle, H. (1981). The interviewing computer: a technology for gathering comprehensive treatment information. *Behavior Research Methods & Instrumentation*, *13*, 607–612.

Bailey, R., Foote, W., & Throckmorton, B. (2000). Human sexual behavior: a comparison of college and internet surveys. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 141–168). San Diego: Academic Press.

Barak, A. (1999). Psychological applications on the Internet: a discipline on the threshold of a new millennium. *Applied and Preventive Psychology*, *8*(4), 231–245.

Barak, A., & English, N. (2001). Prospects and limitations of psychological testing on the Internet. *Journal of Technology in Human Services* (in press).

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice*, *Fall*, 5–12.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: it's not only the scoring. *Educational Measurement: Issues and Practice*, *Winter*, 9–17.

Bennett, R. E., Morley, M., Quardt, D., Rock, D., Singley, M. K., Katz, I. R., & Nhouyvanisvong, A. (1999). Psychometric and cognitive functioning of an under-determined computer-based response type for quantitative reasoning. *Journal of Educational Measurement*, *36*(3), 233–252.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurment*, *34*(2), 162–176.

Birnbaum, M. H. (2000). Decision making in the lab and on the Web. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 3–35). San Diego: Academic Press.

Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 121–140). San Diego: Academic Press.

Buchanan, T., & Smith, J. (1999a). Using the internet for psychological research: personality testing on the World Wide Web. *British Journal of Psychology*, *90*, 125–144.

Buchanan, T., & Smith, J. (1999b). Research on the Internet: validation of a World-Wide Web mediated personality scale. *Behavior Research Methods, Instruments, & Computers*, *31*(4), 565–571.

Burke, M. J. (1993). Computerized psychological testing: impacts on measuring predictor constructs and future job behavior. In Neal Schmitt, Walter C. Borman and Associates, *Personnel selection in organizations*. San Francisco: Jossey-Bass Publishers, pp. 203–239.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kulkich, K., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1998). *Computer analysis of essay content for automated score prediction* (RR-98-15). Princeton, NJ: Educational Testing Service.

Butcher, J., Keller, L., & Bacon, S. (1985). Current developments and future directions in computerized personality assessment. *Journal of Consulting and Clinical Psychology*, *53*, 803–815.

Campbell, K. A., Rohlman, D. S., Storzbach, D., Binder, L. M., Anger, W. K., Kovera, C. A., Davis, K. L., & Grossman, S. J. (1999). Test-retest reliability of psychological and neurobehavioral tests self-administered by computer. *Assessment*, *6*, 21–32.

Carr, A., Ancill, R., Ghosh, A., & Margo, A. (1981). Direct assessment of depression by microcomputer: a feasibility study. *Acta Psychiatrica*, *64*, 415–422.

Carr, A., & Ghosh, A. (1983). Accuracy of behavioral assessment. *British Journal of Psychiatry*, *142*, 66–70.

Choca, J., & Morris, J. (1992). Administering the category test by computer: equivalence of results. *The Clinical Neuropsychologist*, *6*(1), 9–15.

Chua, S. L., Chen, D., & Wong, A. (19999). Computer anxiety and its correlates: a meta-analysis. *Computers in Human Behavior*, *15*, 609–623.

Colby, K. M. (1975). *Artificial paranoia: a computer simulation of paranoid processes*. New York: Pergamon Press.

Colby, K. M. (1981). Modeling a paranoid mind. *The Behavioral & Brain Sciences*, *4*, 515–560.

CommerceNet (2000). CommerceNet [WWW Document]. Available: www.commerce.net/research/stata/april99.html.

Davis, R. (1999). Web-based administration of a personality questionnaire: comparison with traditional methods. *Behavior Research Methods, Instruments, & Computers*, *31*(4), 572–577.

Davis, L., Hoffmann, N., Morse, R., & Luehr, J. (1992). Substance use disorder diagnostic schedule (SUDDS): the equivalence and validity of a computer-administered and interviewer-administered format. *Alcoholism: Clinical & Experimental Research*, *16*(2), 250–254.

Educational Testing Services (2000). *Educational testing services* [Web Document]. Available: http://www.ets.org..

Elwood, D., & Griffin, D. (1972). Individual intelligence testing without the examiner: reliability of an automated method. *Journal of Consulting & Clinical Psychology*, *38*, 9–14.

Erdman, H., Klein, M., & Greist, J. (1985). Direct patient computer interviewing. *Journal of Consulting & Clinical Psychology*, *53*, 760–773.

Evans, L., Tannehill, R., & Martin, S. (1995). Children's reading skills: a comparison of traditional and computerized assessment. *Behavior, Research Methods, Instruments, & Computers*, *27*(2), 162–165.

French, C., & Beaumont, J. (1987). The reaction of psychiatric patients to computerized assessment. *British Journal of Clinical Psychology*, *26*, 267–278.

Georgia Institute of Technology (2000). *GVU's WWW User Surveys* [WWW Document]. Available: www.cc.gatech.edu/gvu/user_surveys.

Green, B., Bock, R., Humphreys, L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360.

Greist, J., Gustafson, D., Strauss, F., Rowse, G., Laughren, T., & Chiles, J. (1973). A computer interview for suicide risk prediction. *American Journal of Psychiatry*, *130*, 1327–1332.

Greist, J., Klein, M., Erdman, H., Bires, J., Bass, S., Machtinger, P., & Kresge, D. (1987, December). Comparison of computer- and interviewer-administered versions of the diagnostic interview schedule. *Hospital & Community Psychiatry*, *38*, 1304–1311.

Hile, M., & Adkins, R. (1997). Do substance abuse and mental health clients prefer automated assessments? *Behavior, Research Methods, Instruments, & Computers*, *29*(2), 146–150.

Hofer, P., & Green, B. (1985). The challenge of competence and creativity in computerized psychological testing. *Journal of Consulting & Clinical Psychology*, *53*, 826–838.

Honaker, L. (1988). The equivalency of computerized and conventional MMPI administration: a critical review. *Clinical Psychology Review*, *8*, 561–577.

Kaye, B., & Johnson, T. (1999). Research methodology: taming the cyber frontier. *Social Science Computer Review*, *17*(3), 323–337.

Krantz, J., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, & Computers*, *29*(2), 264–269.

Krantz, J., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 36–60). San Diego: Academic Press.

Laguna, K., & Babcock, R. L. (1997). Computer anxiety in young and older adults: implications for human-computer interactions in older populations. *Computers in Human Behavior*, *13*(3), 317–326.

Martin, C. L., & Nagao, D. H. (1989). Some effects of computerized interviewing on job applicant responses. *Journal of Applied Psychology*, *74*, 72–80.

Mathisen, K., Evans, F., & Meyers, K. (1987, December). Evaluation of a computerized version of the diagnostic interview schedule. *Hospital & Community Psychiatry*, *38*, 1311–1315.

McCullough, L. (1983). The development of a microcomputer based information system for psychotherapy research. *Problem Oriented Systems & Treatment Post*, *6*(1), 3–4.

Mehta, E., & Sivadas, E. (1995). Comparing response rates and response content in mail versus electronic mail surveys. *Journal of Market Research Society*, *37*, 429–439.

Michalak, E. (1998). The use of the Internet as a research tool: the nature and characteristics of seasonal affective disorder (SAD) amongst a population of users. *Interacting with Computers*, *9*, 349–365.

Millstein, S. (1987). Acceptability and reliability of sensitive information collected via computer interview. *Educational and Psychological Measurement*, *47*, 523–533.

Musch, J., & Reips, U. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61–88). San Diego: Academic Press.

Nadelson, T. (1987, October). The inhuman computer/the too-human psychotherapist. *American Journal of Psychotherapy*, *41*, 489–498.

Nicholson, T., White, J., & Duncan, D. (1998). Drugnet: a pilot study of adult recreational drug use via the WWW. *Substance Abuse*, *19*(3), 109–120.

Pagani, D., & Lombardi, L. (2000). An intercultural examination of facial features communicating surprise. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 169–195). San Diego: Academic Press.

Pasveer, K., & Ellard, J. (1998). The making of a personality inventory: help from the WWW. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 309–313.

Pettit, F. (1999). Exploring the use of the World Wide Web as a psychology data collection tool. *Computers in Human Behavior*, *15*, 67–71.

Pinkerton, G., & Raffoul, P. (1984). Professional colleagues: confronting the attitudes of professionals toward microcomputers. In M. Schwartz (Ed.), *Using computers in clinical practice: psychotherapy and mental health applications* (pp. 61–66). New York: Hawthorne Press.

Pinsoneault, T. (1996). Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota multiphasic personality inventory — 2. *Computers in Human Behavior*, *12*(2), 291–300.

Quintanar, L., Crowell, C., & Pryor, J. (1982). Human-computer interaction: a preliminary social psychological analysis. *Behavior Research Methods & Instrumentation*, *14*, 210–220.

Rasulis, R., Schuldberg, & Murtagh, M. (1996). Computer-administered testing with the Rotter incomplete sentences blank. *Computers in Human Behavior*, *12*(2), 497–513.

Reips, U. (2000). The Web experiment method: advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–120). San Diego: Academic Press.

Richard, D. (2000). Assessing posttraumatic stress symptoms in an international sample via the World Wide Web (unpublished manuscript).

Roid, G., & Gorsuch, R. (1984). Development and clinical use of test-interpretive programs on microcomputers. In M. Schwartz (Ed.), *Using computers in clinical practice: psychotherapy and mental health applications* (pp. 141–149). New York: Hawthorne Press.

Rosen, L., & Weil, V. (1996). Psychologists and technology: a look at the future. *Professional Psychology: Research & Practice*, *27*(6), 635–638.

Rozensky, R., Honor, L., Rasinski, K., Tovian, S., & Herz, G. (1986). Paper-and-pencil versus computer-administered MMPIs: a comparison of patients' attitudes. *Computers in Human Behavior*, *2*, 111–116.

Schmidt, W. (1997). World Wide Web survey research: benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers*, *29*(2), 274–279.

Schmidt, W. (2000). The server side of psychology Web experiments. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 285–310). San Diego: Academic Press.

Slack, W., & Slack, C. (1977). Talking to a computer about emotional problems: a comparative study. *Psychotherapy: Theory, Research, and Practice*, *14*, 156–164.

Smith, M., & Leigh, B. (1997). Virtual subjects: using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 496–505.

Space, L. (1981). The computer as psychometrician. *Behavior Research Methods & Instrumentation*, *13*, 595–606.

Stanton, J. (1998). An empirical assessment of data collection using the Internet. *Personnel Psychology*, *51*, 709–725.

Stones, A., & Perry, D. (1997). Survey questionnaire data on panic attacks gathered using the World Wide Web. *Depression and Anxiety*, *6*, 86–87.

Swoboda, W., Muhlberger, N., Weitkunat, R., & Schneeweiss, S. (1997). Internet surveys by direct mailing. *Social Science Computer Review*, *15*(3), 242–255.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Weiss, D. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*, 774–789.

Weizenbaum, J. (1976). *Computer power and human reason*. San Francisco: Freeman.

Welch, N., & Krantz, J. (1996). The World-Wide-Web as a medium for psychoacoustical demonstrations and experiments: experience and results. *Behavior Research Methods, Instruments, & Computers*, *28*, 192–196.

Wilson, B. (1999). Redressing the anxiety imbalance: computerphobia and educators. *Behaviour & Information Technology*, *18*(6), 445–454.