Journal of Applied Psychology
1999, Vol. 84, No. 5, 754–775

0021-9010/99/$3.00

A Meta-Analytic Study of Social Desirability Distortion in Computer-Administered Questionnaires, Traditional Questionnaires, and Interviews

Wendy L. Richman
University of Illinois at Urbana-Champaign

Sara Kiesler
Carnegie Mellon University

Suzanne Weisband
University of Arizona

Fritz Drasgow
University of Illinois at Urbana-Champaign

A meta-analysis of social desirability distortion compared computer questionnaires with traditional paper-and-pencil questionnaires and face-to-face interviews in 61 studies (1967–1997; 673 effect sizes). Controlling for correlated observations, a near-zero overall effect size was obtained for computer versus paper-and-pencil questionnaires. With moderators, there was less distortion on computerized measures of social desirability responding than on the paper-and-pencil measures, especially when respondents were alone and could backtrack. There was more distortion on the computer on other scales, but distortion was small when respondents were alone, anonymous, and could backtrack. There was less distortion on computerized versions of interviews than on face-to-face interviews. Research is needed on nonlinear patterns of distortion, and on the effects of context and interface on privacy perceptions and on responses to sensitive questions.

As computer and computer-based telecommunications technologies proliferate through society, increasingly they are being used to solicit information from people. Previously existing clinical instruments, personality scales, job attitude scales, cognitive selection tests such as the Graduate Record Examination (GRE), and training inventories are among the many kinds of instruments that have been converted to computerized administration (Mead & Drasgow, 1993). Computer-administered employment, medical and psychiatric intake, consumer preference, and blood donor interviews have been developed to replace both paper-and-pencil and face-to-face interviews, and electronic surveys administered from remote sites already are used to gather personnel, medical, consumer information, and other social science data (Kiesler, Walsh, & Sproull, 1992; Synodinos & Brennan, 1988). Computers offer efficiency advantages over traditional formats, such as reducing transcription errors, and make possible new measurement options such as interactive branching, personalized probes, and provision of explanatory material and on-line help. Should the rapid growth of residential computing continue, psychological assessment using the Internet and the World Wide Web also might become commonplace.

Even as computerized instruments have gained currency in many fields, researchers over the years have regarded their equivalence to traditional formats as somewhat uncertain (e.g., American Institutes for Research [AIR], 1993; Matarazzo, 1983; Potosky & Bobko, 1997; Schuldberg, 1988). In a meta-analysis examining the effect of computerizing cognitive measures such as the GRE, Mead and Drasgow (1993) concluded that mode of administration affects the equivalence of speeded but not power tests. No such meta-analytic review, however, has been conducted for noncognitive instruments. Noncognitive instruments include psychological inventories (e.g., the Minnesota Multiphasic Personality Inventory [MMPI]), attitude scales (e.g., Job Descriptive Index; Smith, Kendall, & Hulin, 1969), behavioral interviews (e.g., intake interviews regarding drug use and abuse), and various scales or subscales of social desirability distortion (e.g., Balanced Inventory of Desirable Responding [BIDR]; Paulhus, 1984). Of particular concern to those who use computerized noncognitive instruments for selection or clinical evaluation is the cross-mode correlation of scores elicited in computerized and traditional modes and the rankings of respondents produced in the two modes. Although some researchers have reported that computerized and traditional instruments elicit highly

Wendy L. Richman and Fritz Drasgow, Department of Psychology, University of Illinois at Urbana-Champaign; Sara Kiesler, the Human–Computer Interaction Institute, Carnegie Mellon University; Suzanne Weisband, Department of Management Information Systems, University of Arizona.

Correspondence concerning this article should be addressed to Wendy L. Richman, William M. Mercer, Inc., 1166 Avenue of the Americas, 23rd Floor, New York, New York 10036. Electronic mail may be sent to Wendy.Richman@us.wmmercer.com.

correlated scores (e.g., Potosky & Bobko, 1997), other studies have shown more modest correlations (e.g., Lushene, O'Neil, & Dunn, 1974; Katz & Dalby, 1981; Vansickle, Kimmel, & Kapes, 1989; White, Clements, & Fowler, 1985; Wilson, Genco, & Yager, 1985).

In that noncognitive instruments typically solicit self-reports of personal information, respondents sometimes slant their responses in a socially desirable direction; their level of social desirability distortion has long been known to vary with the mode of administration (Sudman & Bradburn, 1974). Much of the uncertainty, and the research, about computerized noncognitive instruments has focused on whether the computer mode of administration alters respondents' level of social desirability distortion. Differences in social desirability distortion, in turn, might result in mean differences in scores between computer and traditional instruments. Some investigators have found less social desirability distortion and more candor in responses to a computerized instrument as compared with responses to a traditional instrument (e.g., Evan & Miller, 1969; Kiesler & Sproull, 1986; Martin & Nagao, 1989). Others have found more social desirability and less candor in responses to a computerized instrument as compared with responses to a traditional instrument (Davis & Cowles, 1989; Lautenschlager & Flaherty, 1990; Schuldberg, 1988). Still others reported no differences in responding across administration mode (Booth-Kewley, Edwards, & Rosenfeld, 1992; Erdman, Klein, & Greist, 1983; White et al., 1985). King and Miles's (1995) recent analysis of the factor structure of four computerized and paper-and-pencil instruments suggests that traditional and computerized modes of administration have similar latent structures; yet their respondents received significantly higher scores on a measure of social desirability distortion in the paper-and-pencil version than in the computer version.

On the basis of the long-standing interest of researchers in the equivalence of computerized and traditional noncognitive instruments, the absence of a previous meta-analysis of this literature, and inconsistent reports of social desirability response effects, we undertook this study. The study had two main purposes. Its first purpose was to evaluate the literature comparing computer noncognitive instruments with traditional instruments to assess the extent to which mean differences in scores exist. We also made a separate examination of the effects of computer versus traditional modes of administration on direct measures of social desirability distortion. Our second purpose was to examine the conditions under which computer instruments and traditional instruments yield nonequivalent results. For example, if computerized and traditional modes of administration differ in the degree to which they support perceptions of anonymity, and if anonymity reduces social desirability distortion, then anonymous or identified data collection should affect mean score differences across modes. Identifying such conditions could inform the design of new computer instruments whose presentation features or response alternatives may differ from traditional formats. To these ends, we conducted a meta-analysis of the published literature that has accumulated for the last 30 years comparing computer and traditional noncognitive questionnaires, psychological tests, and interviews.

Social Desirability Distortion

Social desirability distortion refers to the tendency by respondents, under some conditions and modes of administration, to answer questions in a more socially desirable direction than they would under other conditions or modes of administration. Other terms used in this literature are *response bias* (Rezmovic, 1977), *socially desirable responding* (Zerbe & Paulhus, 1987), *response distortion* (Potosky & Bobko, 1997) and *overreporting* (Turner et al., 1998). Social desirability distortion can result from quite different processes. For example, unintentional distortion can occur through carelessness and disinterest, mood changes, changes in the depth of cognitive processing about the self, or overconfidence (e.g., Dunning, Griffin, Milojkovic, & Ross, 1990). Paulhus (1984) distinguished between unintentional self-deception and intentional impression management. Intentional impression management might involve strategically "faking bad" to obtain a resource, such as disability compensation, or sympathy, or it could involve "faking good" to make a good impression or to hide sensitive personal information. Most discussions in the literature on computer instruments have referred to the latter, that is, whether respondents using a computer instead of a traditional instrument will be more or less likely to purposely distort their responses in a socially desirable direction. Although the existence and impact of intentional distortion on prediction or evaluation is open to some debate (Douglas, McDaniel, & Snell, 1996; Ones, Viswesvaran, & Reiss, 1996), evidence suggests that the usefulness of noncognitive measures is eroded when respondents intentionally bias their answers in a socially desirable direction (Snell & McDaniel, 1998; Zickar & Robie, 1998).

Several arguments have been advanced concerning the specific role of social desirability distortion in computerized instruments. One argument, dating from the 1960s when computers were first used for measurement (Smith, 1963), is that respondents using a computer as compared with traditional instruments may feel anonymous, private, or free of social pressure and the evaluative "test" situation, and hence may be less prone to give socially desirable answers (or more prone to be candid). One of the first applications of a computerized instrument was a medical intake interview on allergy problems, in which the investigators reported that the computer interview elicited more complete information than medical charts (Slack, Hicks, Reed, & Van Cura,

1966). In 1969, Evan and Miller reported that respondents were more open to the computer for "highly personal and possibly disturbing" content areas (p. 216). Greist and his colleagues, who developed several psychiatric interviewing systems, reported that "computer administration seems to make patients more comfortable and candid than does a human interviewer" especially in "interviews for sexual dysfunction, and alcohol and drug use" (Greist & Klein, 1981, p. 767). Sproull and Kiesler (1986) argued that computerized instruments (of that era) typically lacked social context cues such as visual reminders of the test format and organizational identifiers. The absence of familiar test cues might make computer instruments seem evaluatively neutral, not only in comparison to a real interviewer, but also in comparison to traditional printed questionnaires or surveys.

Research on computer-mediated communication (e.g., Dubrovsky, Kiesler, & Sethna, 1991; Short, Williams, & Christie, 1976; Sproull & Kiesler, 1986) indicates that people often experience a feeling of privacy or anonymity while communicating through the computer. Some investigators have speculated that computer administration might increase perceptions of the measurement situation as private or anonymous (e.g., Evan & Miller, 1969; Griffith & Northcraft, 1994; Kantor, 1991; Lucas, Mullin, Luna, & McInroy, 1977; Robinson & Walters, 1986; Turner et al., 1998). A large literature shows anonymity reduces social desirability distortion and increases self-disclosure (Cannavale, Scarr, & Pepitone, 1970; Diener, Fraser, Beaman, & Kelem, 1976; Lindskold & Finch, 1982; Mathes & Guest, 1976; Rosenfeld, Giacalone, & Riordan, 1992; Singer, Brush, & Lublin, 1965; White, 1977; Zerbe & Paulhus, 1987). Instructions that responses to questions will be anonymous should serve to reassure respondents, and this instruction might be especially effective when respondents are using a computer instrument that does not request identifying information and conveys few social context "test" cues. Also, if respondents type responses that seem to disappear into the computer, they may feel more anonymous than they do taking traditional tests, in which there is a concrete reminder of the evaluation, such as a printed questionnaire or interviewer (Sproull & Kiesler, 1991).

Other researchers have argued that computer administration in some cases could reduce feelings of privacy and anonymity (e.g., Rosenfeld, Booth-Kewley, Edwards, & Thomas, 1996; Yates, Wagner, & Suprenant, 1997). Rosenfeld et al. (1996) argued that a computer can seem threatening rather than safe when respondents know their responses will be identified, verified, and stored in a database. In their study of the "big brother syndrome" in a Navy boot camp, recruits gave responses indicating higher levels of impression management when they were identified and used a networked computer, a finding that supports the authors' claim. Yates et al. (1997) displayed a computerized questionnaire on an overhead projector and asked groups of undergraduates to answer sensitive questions using keypads at their seats. The students perceived this computer condition to be much less anonymous than the paper-and-pencil comparison condition.

A different line of argument concerns the effects on social desirability distortion of technological advances in computer instruments themselves. When computers were first used in assessment, items were often displayed on terminals as green phosphorous uneditable screens of text, and respondents typically could not go back to review their responses. Researchers have argued that if respondents are constrained by the computer in the ways they can look at items and answer (e.g., by not being able to skip items, not having the option of answering "don't know," or not being allowed to backtrack), they may feel self-conscious or wary (e.g., Spray, Ackerman, Reckase, & Carlson, 1989). As a result, respondents may give more socially desirable responses on the computer instrument. Newer hardware and software have made it possible to improve the interface, and now give investigators many options for presentation of items. For instance, computer forms can resemble printed "interviews" (e.g., Intuit's MacInTax software), or they can look just like standardized printed questionnaires, complete with official trademarks or logos and fill-in-the-blank items. One might argue on this basis that there can be no overall effect of computerization. The effect will depend on how the interface makes respondents feel; the more a computer instrument resembles a traditional instrument, the more the two instruments should produce similar responses.

Meta-Analysis

Because a vast majority of studies have compared mean differences between the scores obtained when the mode of administration was a computer instrument as compared with a paper-and-pencil instrument or face-to-face interview, our meta-analysis first addressed whether mean scores in noncognitive assessments have in fact differed across administration mode and the direction of social desirability distortion implied by these differences. Because the literature suggests that mean differences vary with type of instrument and features of the administration context and interface, we did not make a prediction as to the overall effect of administration mode across all studies.

Hypotheses

We used hierarchical regression to test hypotheses about the conditions under which computer instruments would generate more or less social desirability distortion than comparable traditional instruments. We pursued this exploration using a statistical analysis that allows clustered (i.e., correlated) observations so as to fully use all interpretable effect sizes reported in the literature.

Mode comparison. The nature of the cross-mode comparison has received practically no discussion in the literature we reviewed. Yet today's computer instruments have much more in common with paper-and-pencil instruments than with face-to-face interviews, particularly in respect to factors that should affect social desirability distortion. In the literature we reviewed, both computer and paper-and-pencil instruments displayed printed items, were self-administered (respondents either typed into a computer or wrote on a questionnaire), could be made anonymous or identified, and could be administered when respondents were completely alone or with others. Many (but not all) investigators used the same levels of anonymity and presence of others in both the computer and paper-and-pencil conditions of their studies.

By contrast, comparisons of computer "interviews" with face-to-face interviews nearly always have confounded mode of administration with other important variables, particularly self-administration, anonymity, and the presence of others. In some studies self-administration has been varied independently or controlled (e.g., Robinson & West, 1992), but the interviewer's presence in the face-to-face condition would have reduced perceptions of anonymity, and the interviewer's appearance, intonation, hesitations, facial expressions, or gestures would have conveyed more social expectations and social pressure than the computer instrument would have conveyed. It has been suggested that face-to-face interviews are more likely to elicit social desirability distortion than are paper-and-pencil questionnaires (Sudman & Bradburn, 1974). Face-to-face interviews appear to differ from computer questionnaires in much the same way.

> *Hypothesis 1:* The difference between a computer instrument and a traditional instrument will be greater in magnitude, and in the direction of less social desirability distortion on the computer, when the traditional instrument is a face-to-face interview rather than a paper-and-pencil questionnaire.

Measures of social desirability distortion. As noted above, researchers have posited that social desirability distortion is the principle cause of mean score differences between computer and traditional instruments. To examine this idea, researchers comparing computerized and traditional modes of administration have used measures of social desirability distortion as dependent variables. Measures of social desirability distortion are operationalizations of respondents' response bias due to their desire to appear socially desirable. These measures include the Social Desirability Scale (Edwards, 1957), the K (Defensiveness) and L (Lie) scales of the MMPI, the Crowne–Marlowe Need for Approval scale (Crowne & Marlowe, 1964), and the BIDR, which attempts to measure innocent positive self-deception as well as strategic positive impression management (Paulhus, 1984). In a few studies, researchers' dependent measure of distortion was the difference between self-reported and objective data about participants. For instance, Martin and Nagao (1989) compared college students' self-reported grade point averages and college entrance examination test scores with their actual scores. Waterton and Duffy (1984) compared self-reported alcohol consumption with actual alcohol sales figures.

Researchers also have examined indirect evidence of social desirability distortion in many psychological, medical, personality, and employment scales, inferring the degree of distortion from the extent to which scores obtained were in the direction of normality or social desirability. Less distortion might be inferred from more negative scores on measures of medical, educational, or psychological problems. For example, elevated scores on an instrument measuring anxiety would be considered indicative of low social desirability distortion. We reasoned that instruments specifically designed to measure social desirability distortion have face validity as measures of social desirability response bias, and are more reliable and direct indicators of distortion than instruments used to measure diverse other traits, syndromes, attitudes, and behavior. We therefore decided to split the analysis and evaluate results for the two kinds of instruments separately. We first evaluated the results of comparisons of computer and paper-and-pencil measures of social desirability distortion. Then we evaluated the results of comparisons of computer and paper-and-pencil measures of all other traits, attitudes, and behaviors. Although we did not formulate a testable hypothesis, we expected differences in social desirability distortion to be evident in studies using measures of distortion, whereas these differences might not be evident in studies using other measures.

Perceptions of anonymity, neutrality, and privacy. Researchers have argued that the evaluation situation carries more neutrality and anonymity when the mode of administration is by computer, in part because of the absence of social context information in the computer and the perception that responses disappear into the computer screen. This process ought to be moderated by context factors such as anonymity or identity instructions, whether or not respondents are alone (interacting only with the computer), and whether or not the computer's information processing capabilities arouse privacy concerns. Typing into a computer may reinforce anonymity instructions, resulting in less social desirability distortion with a computerized instrument than a paper-and-pencil instrument.

> *Hypothesis 2:* When respondents are assured anonymity, less social desirability distortion will occur with a computer instrument than with a paper-and-pencil instrument.

The presence of an interviewer, test administrator, experimenter, or other respondents in the testing situation might obviate claims of anonymity and remind respondents of the

evaluative test nature of the situation, whereas being alone would reinforce the sense of neutrality and privacy.

Hypothesis 3: When respondents are alone while responding to items on a computer, less social desirability distortion will occur with a computer instrument than with a paper-and-pencil instrument.

Ease of response and other interface attributes. Perhaps because of technical limitations of their computers or software, some investigators using computer instruments have prevented respondents from reviewing, skipping, or changing their responses as they could with paper questionnaires. Such constraints may lead respondents to worry about making mistakes, to process their responses more thoughtfully, or to feel a lack of control, and then to increase their social desirability distortion (e.g., Spray et al., 1989). For example, Lautenschlager and Flaherty (1990) argued that without being able to check previous responses "regarding how much one has already distorted responses in certain situations, one may be more prone to further distort any given item" (p. 313).

Hypothesis 4: Not being able to skip items (or to answer "not applicable" or "don't know") and not being able to backtrack and edit answers will lead to greater social desirability distortion on computer instruments relative to traditional instruments that typically include these features.

Sensitive information. Catania, Gibson, Chitwood, & Coates (1990) have proposed that socially desirable distortion is likely to increase when the items are of a highly sensitive nature, dealing, for example, with sexuality or illegal behavior. When asked to answer sensitive information, respondents may be wary of responding with candor because they may have something to hide. A mode of administration that seems to protect the person's identity or anonymity should have a greater impact when the items are sensitive than when they are not (Catania et al., 1990).

Hypothesis 5: When instruments solicit sensitive, personal, or otherwise risky information as compared with impersonal information, less social desirability will occur with the computer instrument than with the traditional instrument.

Method

Literature Review

We manually searched the literature from 1967 to 1997 using *Psychological Abstracts,* Social Citations Index, Social Sciences Citation Index, and reference lists of all the articles we located. We used the keywords and keyword phrases that included *computer forms, instruments, questionnaires,* or *interviews; measurement, distortion, disclosure, anonymity, social desirability, socially desirable, response effects,* and *response bias.* We ultimately obtained studies from the social sciences, computer sciences, and medical literature and communicated with colleagues to search for additional relevant research. Authors of studies found in the initial literature search were contacted to obtain additional information and sources. We also relied on our own knowledge of the literature on social aspects of computer technology to find articles and books that might contain references to studies of social desirability distortion and self-disclosure. We made every effort to find all eligible manuscripts; however, undoubtedly we missed some studies.

We included or excluded studies in our meta-analysis according to the following rules: A study had to be a published investigation of computerized questionnaires, noncognitive tests, or structured interviews; we did not include studies of email communication or group discussion using computers (for a review of that literature, see Kiesler & Sproull, 1992). We excluded studies of cognitive ability or achievement such as the GRE (for a review of the effect of computer administration of cognitive tests, see Mead & Drasgow, 1993). To be included in the analysis, studies had to have included a comparison or control group (e.g., a condition in which respondents completed a paper-and-pencil instrument); studies that lacked a comparison group or seriously violated the usual standards of experimental design (e.g., compared data from two different sources at two different times) were excluded. Also, the studies that we included had to be described sufficiently, with the statistics necessary to compute an effect size, although in several cases we were able to include studies when the authors provided us with the needed information. We included studies in which the computerized instrument displayed items on a desktop or laptop computer, terminal, or workstation and respondents typed or keyed in their answers. We did not include a few recent studies of computer-based systems in which items were administered using a group projection screen, or audio or telephone systems (Tourangeau & Smith, 1996; Turner et al., 1998; Yates, Wagner, & Suprenant, 1997).

We excluded unpublished studies from our analysis because we were able to locate very few that would have qualified by our criteria. In this domain of research (equivalence of forms of administration), journals have published well-crafted studies in which the authors found no differences between the computer instrument and the traditional instrument. Thus, we believe our exclusion of unpublished manuscripts is unlikely to have created a significant "file drawer" problem.

With the above restrictions, the final sample included a collection of 673 effect sizes from 61 studies spanning almost 30 years of research. Table 1 lists the studies used in the analysis.

Moderators

On the basis of our hypotheses, we coded the following variables for each study: cross-mode comparison conditions (computer vs. paper-and-pencil questionnaire or computer vs. face-to-face interview); type of noncognitive assessment (personality inventory, attitude scale, symptom checklist, or social desirability measure such as the BIDR); anonymity instructions (anonymous or identified); presence of one or more others during administration of the instruments (alone or not alone); whether or not the instrument requested highly sensitive personal information; computer response options (skipping and backtracking options available or not available). When the comparison condition was a face-to-face interview, anonymity and presence of others were coded for the computer condition only.

We also examined various study characteristics. The study char-

acteristics included study design (between or within groups); type of subject population (student, substance abuse, psychiatric, or other adult); percentage of women in the sample; mean age of the sample; year of publication; and reliability of the dependent measures.

Operationalization of the moderators was straightforward except in the categorization of measures as sensitive. A measure was coded as sensitive if it asked for personal information not normally discussed among casual acquaintances, such as information about the respondent's finances, criminal record, sexual behavior, or use of drugs. Psychiatric tests and behavior checklists regarding sexual practices or drug usage were coded as sensitive. The first and second authors independently coded the moderator variables. The percentage of agreement was 89%. In cases of disagreement, additional information was gathered from the authors of the studies and any remaining differences were resolved through discussion.

Procedure

Meta-analysis is currently receiving a great deal of attention from statisticians. Applied researchers may be most familiar with the approach to meta-analysis called validity generalization (Hunter, Schmidt, & Jackson, 1982) in which the observed variance of a set of correlations is compared to the variance expected on the basis of artifacts (sampling variability, restriction of range, criterion unreliability). Validity generalization is just one analysis from a much larger collection of methods; indeed, Hedges and Olkin's (1985) book documents numerous methods.

Our approach to meta-analysis used multiple regression to examine the effects of various moderators. The dependent variable was the effect size, *d*, computed as:

$$d = \frac{M_c - M_t}{s},$$

where M_c and M_t denote scale means for the computerized and traditional comparison groups (paper-and-pencil or face-to-face format), respectively, and s is the pooled within-group standard deviation (Hedges & Olkin, 1985). As noted previously, most personality scales, symptom checklists, and attitude measures were not developed as measures of social desirability distortion. We made certain assumptions about the relationship between higher scores on these scales and social desirability distortion. If most respondents attempting to be socially desirable want to appear normal and healthy (fake good), then scores reflecting less illness, fewer symptoms, or lower abnormality (e.g., lower MMPI scores for depression and anxiety) should reflect more social desirability response distortion. In this vein, we recoded all scales so that higher scores always referred to greater social desirability distortion (e.g., attempting to look good, failing to discuss personal problems, etc.). In our analyses, effect sizes are positive when respondents apparently engaged in more social desirability distortion on the computer instrument compared with the traditional instrument and negative when respondents apparently engaged in less social desirability distortion on the computer instrument compared to the traditional instrument.

We performed a series of regression analyses in an attempt to find a parsimonious model that accurately predicted the effect size statistic as a function of the comparison modes, type of assessment, computer software interface, presence of others, anonymity, and other study characteristics that we coded. Such a model could then be used to interpret the consequences of potential moderators.

Two major difficulties were encountered in our analyses. First, some of the independent variables were highly correlated, which created problems frequently observed in the presence of multicollinearity (e.g., large standard errors for regression coefficients). For example, the age of the subject population and backtracking were highly correlated ($r = .73$), possibly because of differences in the types of computers used in schools versus other settings. When two moderators were highly correlated, we did not enter both into the regression equation. As described below, only subsets of variables were entered in order to minimize multicollinearity effects and simplify interpretation.

The second problem resulted from the fact that many studies reported more than a single effect size. When multiple effect sizes were reported for a given study, it seemed unlikely that treating each effect size as a separate case for the regression analyses would satisfy the standard statistical assumption that one's observations are independent. In past meta-analyses, researchers have considered (a) randomly picking one effect size per study, (b) averaging across the effect sizes reported for a single study, or (c) ignoring this violation of a critical statistical assumption. The first two options have the advantage of satisfying the assumption that the cases used for the regression analyses are independent; unfortunately, these approaches discard substantial amounts of information. The latter approach uses all the available data for analysis, but creates questions about the accuracy of the hypothesis tests.

In our meta-analysis, we used all the data available, but explicitly accounted for the "correlated error" of multiple effect sizes from a single study. Specifically, multiple effect sizes from a single study were treated as observations from a clustered sample (i.e., nonindependent observations; see Cochran, 1977, for statistical details). The SUDAAN computer program (Shah, Barnwell, & Bieler, 1995) accounts for clustered observations by increasing the magnitude of the standard error for parameter estimates to the degree that observations within clusters provide redundant information. Point estimates of parameters, however, are unaffected by nonindependence of observations (Shah et al., 1995). We used multiple effect sizes from individual studies (when they were reported) in regression analyses that explicitly corrected for statistical dependencies among observations.

We examined the design effect, a statistic that indexes the degree of dependence in clustered data, to determine the extent to which effect sizes within studies were not independent (and hence, the degree to which it was necessary to correct the standard errors of regression coefficients). The design effect is defined as:

$$deff = \frac{\text{variance for estimate from actual sampling design}}{\text{variance for estimate from a simple random sample}}.$$

Thus, the design effect is less than 1 when a sampling plan (for example, a stratified random sample) is more efficient than a simple random sample. In our case, with clustered observations, we were interested in the degree to which the design effect would be greater than 1; this would inform us about the degree to which dependencies of within-study effect sizes reduced the amount of information provided by the data.

(*text continues on p. 763*)

Table 1
Summary of Meta-Analysis Study Characteristics

Study	Design[a]	Mean computer *N*	Mean control *N*	Female % in sample	*M* age (years)	Sample	Categorical variables[b]	Measures[c]	Mean effect size (*d*)[d]	No. of effect sizes to compute *M*
Computer versus paper-and-pencil instruments										
Biskin & Kolotkin (1977, Study 1)	B	37	45	.00	19.5	Undergraduates	1/2/2/1	MMPI	0.23	14
Biskin & Kolotkin (1977, Study 2)	B	18	21	.00	19.5	Undergraduates	1/2/2/1	MMPI	0.17	14
Booth-Kewley et al. (1992)	B	40	42	.00	20.0	Navy recruits	3/1/1/3	**Impress Mgmt (BIDR)**	0.08	4
								Self-Deception (BIDR)	−0.07	4
Davis & Cowles (1989)	B	72	75	.52	19.5	Undergraduates	1/2/1/1	**Crowne–Marlowe**	−0.24	2
								Locus of Control (Rotter)	0.30	2
								Anxiety scale	0.26	2
								Eysenck Personality Inventory	−0.03	6
Erdman et al. (1983)	W	133	133	—	—	HS students	2/2/–/2	Cigarette/Drug use	−0.24	1
Evan & Miller (1969)	B	30	30	.00	19.5	Undergraduates	2/2/–/–	Manifest Anxiety (MMPI)	0.00	1
								Lie Scale (MMPI)	0.31	1
								Allport–Vernon values	0.89	1
								Srole Anomie	0.00	1
								Factual information	−0.48	1
								Impersonal items	0.64	1
								Pool personal items	−0.48	1
								Pool impersonal items	0.03	1
Finegan & Allen (1994, Study 1)	B	31	32	.49	19.5	Undergraduates	2/1/2/1	**Social Desirability**	0.51	1
								Achievement motivation	0.61	1
								Attitude scale	0.25	1
								Anxiety scale	−0.48	1
Finegan & Allen (1994, Study 2)	W	40	40	.80	19.5	Undergraduates	2/1/2/1	**Social Desirability**	0.00	1
								Attitude scale	0.44	3
Finegan & Allen (1994, Study 3)	B	36	31	.55	19.5	Undergraduates	2/1/2/1	**Social Desirability**	0.00	1
								Crowne-Marlowe	0.00	1
								Self-Deception	0.00	1
French & Beaumont (1989)	W	124	203	.37	38.4	Psychiatric patients	1/–/2/2	Eysenck Personality Inventory	0.04	8
George et al. (1992)	B	48	49	.54	19.0	Undergraduate	–/1/–/2	Beck Depression Inventory	−0.44	1
								State–Trait Anxiety Inventory	−0.27	2
Greist et al. (1975)	B	50	50	—	—	Substance abuse patients	2/2/–/2	Homosexuality	−0.47	1
								Murder wish	−0.30	1
								Premarital sex	−0.29	1
Hart & Goldstein (1985)	B	20	10	.00	43.6	Psychiatric patients	2/–/2/–	MMPI	0.28	28
Hinkle et al. (1991)	W	22	22	.82	31.5	Psychiatric patients	–/–/–/2	Personal problem checklist	−0.14	1
Honaker et al. (1988)	B	20	20	.50	30.9	Adults	–/–/2/2	MMPI	−0.08	80
Honaker et al. (1988)	W	20	20	.50	30.9	Adults	–/–/2/2	MMPI	0.03	80
Kantor (1991)	B	92	84	.20	45.0	Adults	–/1/2/–	Job Description Index	−0.05	5
Katz & Dalby (1981)	W	18	18	.50	28.3	Psychiatric patients	2/1/1/1	Eysenck Personality Inventory	0.63	3
Kiesler & Sproull (1986)	B	49	51	.16	19.5	Undergraduates	2/2/–/–	**Crowne–Marlowe**	0.39	1
								Personal questions	−0.55	1
Kiesler & Sproull (1986)	B	13	20	.16	19.5	Undergraduates	2/2/–/–	**Crowne–Marlowe**	0.64	1
King & Miles (1995)	B	483	391	.46	21.0	Undergraduates	1/1/1/2	**Self-Deception (BIDR)**	0.02	1
								Impress Mgmt (BIDR)	−0.18	1
								Mach V scale	−0.10	1
								Equity Sensitivity	−0.15	1
								Self-Esteem (Rosenberg)	−0.05	1
Koson et al. (1970)	B	16	16	.50	20.0	Undergraduates	2/2/–/–	Threat (MMPI)	−0.27	1
								K Scale	0.06	1
Lambert et al. (1987)	W	21	21	.00	42.4	Medical patients	1/1/1/2	MMPI	−0.01	28
Lankford et al. (1994)	B	65	66	.49	19.5	Undergraduates	–/–/–/–	Beck Depression Inventory	−0.16	1
								Purpose of Life (Beck)	0.30	1
Lautenschlager et al. (1990)	B	42	39	.70	18.8	Undergraduates	3/–/1/1	**Impress Mgmt (BIDR)**	0.37	2
								Self-Deception (BIDR)	0.14	2
Liefeld (1988)	B	239	261	—	—	Adults	–/1/–/–	Attitude scale	0.11	8

Table 1 (*continued*)

Study	Design[a]	Mean computer *N*	Mean control *N*	Female % in sample	*M* age (years)	Sample	Categorical variables[b]	Measures[c]	Mean effect size (*d*)[d]	No. of effect sizes to compute *M*
Locke & Gilbert (1995)	B	54	54	.50	19.8	Undergraduates	2/2/2/1	**MMPI F scale**	−0.29	1
								Drinking habits	0.65	9
Lukin et al. (1985)	W	66	66	.67	19.5	Undergraduates	–/2/–/2	Reactance Scale	0.00	1
								Trait Anxiety	0.04	1
								Beck Depression Inventory	0.00	1
Lushene et al. (1974)	W	31	31	1.0	19.5	Undergraduates	2/1/2/1	MMPI	0.17	26
Martin & Nagao (1989)	B	25	27	.34	19.5	Undergraduates	1/2/1/2	Locus of Control (Rotter)	−1.81	1
Martin & Nagao (1989)	B	19	23	.34	19.5	Undergraduates	1/2/1/2	**SAT bias**	0.62	1
								GPA bias	0.00	1
Millstein (1987)	B	33	43	1.0	16.9	Medical patients	2/2/2/1	Sexual behavior	0.00	1
								Substance abuse	0.16	1
								Gynecologic symptoms	0.12	1
								Nongynecologic symptoms	0.00	1
								Positive affect	0.82	4
								Negative affect	−0.31	5
Potosky & Bobko (1997)	W	176	176	.55	19.5	Undergraduates	1/1/2/2	**Impress Mgmt (BIDR)**	0.06	1
								Self-Deception (BIDR)	−0.03	1
								Experiences and Attitudes	0.00	9
Rezmovic (1977)	W	49	49	.29	19.5	Undergraduates	2/2/–/–	Locus of Control	0.08	2
								Crowne–Marlowe	−0.03	2
Ridgway et al. (1982)	W	27	27	.00	26.8	Nurses	2/–/–/–	Eysenck Personality Inventory	0.05	6
Robinson & West (1992)	B	37	32	.52	27.0	Psychiatric patients	1/–/1/1	No. of symptoms	−0.52	1
								No. of previous visits	−0.31	1
								No. sex partners	−0.21	1
Rosenfeld et al. (1989)		148	148	.00	—	Adults	2/–/–/–	Decision-making survey	0.21	1
Rosenfeld et al. (1991)		36	36	—	19.5	Undergraduates	1/–/–/–	Job Description Index	−0.06	1
Rosenfeld et al. (1996)		42	42	.00	19.1	Adults	3/1/1/1	**Impress Mgmt (BIDR)**	0.29	2
								Self-Deception (BIDR)	0.39	2
Rozensky et al. (1986)	B	86	86	.28	35.1	Psychiatric patients	1/1/–/–	Computer Attitude Survey	−0.20	8
Schuldberg (1988)	W	150	150	.44	19.5	Undergraduates	2/1/–/–	MMPI	0.07	14
Scissons (1976)	B	20	18	.00	19.5	Undergraduates	–/–/2/2	California Psychological Inventory	−0.70	11
Skinner & Allen (1983)	B	50	50	.24	28.0	Medical patients	1/2/–/–	Michigan Alcoholism Screen	0.00	1
Synodinos et al. (1994)	B	265	274	—	40.0	Adults	–/–/–/–	Complaints	−0.09	12
Watson et al. (1990)	W	200	200	.00	38.1	Psychiatric patients	1/–/1/–	MMPI	0.02	13
White et al. (1985)	B	25	25	.55	18.5	Undergraduates	1/–/2/1	MMPI	0.01	28
White et al. (1985)	W	25	25	.55	18.5	Undergraduates	1/–/2/1	MMPI	0.04	56
Whitener & Klein (1995)	B	20	20	.51	19.5	Undergraduates	1/1/1/3	Need for Achievement	0.72	4
								Locus of Control (Rotter)	0.64	4
								Self-Esteem (Rosenberg)	0.42	4
								Impress Mgmt	0.77	4
Wilson et al. (1985)	W	64	64	1.0	19.4	Undergraduates	–/–/–/–	Test Attitude Battery	0.15	6
					Computer versus face-to-face interviews					
Alemi & Higley (1995)	B	35	45	.47	36.6	Adults	1/2/2/2	Risk Factors	0.46	1
Angle et al. (1979)	W	55	55	—	—	Psychiatric patients	1/–/–/–	Mental health problems	−2.43	1
								Specific problems	−2.22	1
Barron et al. (1987)	B	19	19	—	20.0	Undergraduates	–/2/1/–	Client Expectancy Survey	0.06	11
Canoune & Leyhe (1985)	W	52	52	.50	19.5	Undergraduates	–/2/–/–	Interpersonal values	0.11	3
Carr & Ghosh (1983)	W	26	26	—	—	Psychiatric patients	1/2/–/–	Phobia questionnaire	−3.59	1
								Symptom checklist	−1.85	1

(*table continues*)

Table 1 (*continued*)

Study	Design[a]	Mean computer *N*	Mean control *N*	Female % in sample	*M* age (years)	Sample	Categorical variables[b]	Measures[c]	Mean effect size (*d*)[d]	No. of effect sizes to compute *M*
Carr et al. (1983)	W	37	37	.41	37.0	Psychiatric patients	1/2/1/1	Masturbation	−0.91	1
								Criminal record	−0.82	1
								Impotence	−0.69	1
								Alcohol/drug abuse	−0.76	1
								Fired from job	−0.63	1
								Suicide attempt	−0.63	1
Davis et al. (1992)	W	100	100	.35	41.4	Psychiatric patients	1/–/–/–	Drug symptoms	0.24	1
Erdman et al. (1992)	W	78	78	.61	36.0	Psychiatric patients	1/2/–/2	NIMH Diagnostic	0.07	1
Farrell et al. (1987)	W	103	103	.70	25.0	Psychiatric patients	1/2/2/2	No. of complaints	−0.82	1
Ferriter (1993)	B	10	10	—	38.0	Psychiatric patients	1/–/–/–	No. of extreme responses	−0.20	1
Greist et al. (1987)	W	150	150	—	37.6	Psychiatric patients	1/2/–/2	Diagnostic Interview	−0.08	1
Kobak et al. (1990)	W	32	32	.56	31.4	Adults	1/2/1/2	Hamilton Depression Scale	0.09	3
Kobak et al. (1993)	W	97	97	.51	37.0	Adults	1/2/1/2	Hamilton Anxiety Scale	−0.20	3
Koson et al. (1970)	B	16	16	.50	20.0	Undergraduates	2/2/–/–	Threat (MMPI)	−0.38	1
								K Scale	0.29	1
Levine et al. (1989)	W	102	102	.61	32.5	Psychiatric patients	–/–/–/–	Suicide prediction	−2.81	1
Liefeld (1988)	B	239	288	—	—	Adults	–/1/–/–	Attitude Scale	0.05	8
Locke & Gilbert (1995)	B	54	54	.50	19.8	Undergraduates	2/2/2/1	**MMPI F scale**	0.59	1
								Drinking Habits	0.44	8
Locke et al. (1992)	W	272	272	.51	40.0	Adults	2/–/–/–	HIV Risk	−0.30	2
Lucas et al. (1977)	W	36	36	.00	—	Substance abuse patients	1/–/2/–	Alcohol consumption	−0.69	1
Martin & Nagao (1989)	B	25	27	.34	19.5	Undergraduates	1/2/1/2	Locus of Control	−1.92	1
Martin & Nagao (1989)	B	19	36	.34	19.5	Undergraduates	1/2/1/2	**SAT bias**	0.54	1
								GPA bias	0.61	1
Millstein (1987)	B	33	32	1.0	16.9	Medical patients	2/2/2/1	Sexual behavior	−0.21	1
								Substance abuse	−0.24	1
								Gynecologic symptoms	−0.18	1
								Nongynecologic symptoms	0.18	1
								Positive affect	1.03	4
								Negative affect	−0.39	5
Robinson & West (1992)	W	37	37	.52	27.0	Psychiatric patients	1/–/1/1	No. of symptoms	−1.12	1
								No. of previous visits	−0.26	1
								No. of sex partners	−0.33	1
R. Rosenfeld et al. (1992)	W	24	24	.50	35.8	Psychiatric patients	1/1/1/2	Obsessive Scale	−0.12	2
								Compulsive Scale	−0.11	2
R. Rosenfeld et al. (1992)	W	23	23	.50	23.7	Adults	1/1/1/2	Obsessive Scale	−0.57	1
								Compulsive Scale	−0.83	1
Skinner & Allen (1983)	B	50	50	.24	28.0	Medical patients	1/2/–/–	Michigan Alcoholism Screen	−0.10	1
Sproull (1986)	W	48	48	.38	34.0	Adults	1/2/–/–	No. of extreme responses	0.47	1
Waterton & Duffy (1984)	B	145	175	.00	—	Adults	–/2/1/1	**Alcohol consumption**	−0.19	3

Note. Dashes indicate missing data. MMPI = Minnesota Multiphasic Personality Inventory; Mgmt = management; BIDR = Balanced Inventory of Desirable Responding; HS = high school; SAT = Scholastic Assessment Test; GPA = grade point average; NIMH = National Institute of Mental Health. [a] Within-subjects (W) or between-subjects (B) design. [b] The first variable is whether the participant's responses on the computer were anonymous (1 = identified, 2 = anonymous, 3 = manipulated anonymity); the second variable is whether the participant answered questions on the computer without the presence of others (1 = not alone, 2 = alone, 3 = manipulated presence of others); the third variable is whether the participant could skip questions on the computer (1 = skip option not available, 2 = skip option available, 3 = manipulated availability of the skip option); and the fourth variable is whether the participant could backtrack to previous questions on the computer (1 = backtracking not available, 2 = backtracking available, 3 = manipulated availability of the backtracking option). [c] Bolded entries are measures of social desirability distortion. [d] Effect sizes were positive when there was more social desirability response distortion in the computer condition than in the comparison condition and negative when there was less social desirability response distortion in the computer condition than in the comparison condition.

In sum, our main data-analytic method was multiple regression. The regression coefficients we report can be obtained from any standard statistical software (e.g., SPSS or SAS). The principal difference between our approach and that of others is that the standard errors of the regression coefficients were corrected for nonindependence of some observations.

Results

Descriptive Statistics

Table 1 presents mean effect sizes from each study along with a summary of study attributes and an abbreviated description of the measures used in each study. In the first analysis, the SUDAAN program was used to compute the mean and standard error of the entire set of 673 effect sizes across 61 studies. The mean effect size for computer versus all traditional instruments was .02 (standard error computed with SUDAAN = .04), indicating that computerized administration had no overall effect on social desirability distortion.

Preliminary examination of all 673 effect sizes revealed that several of the study characteristic moderators were not significant in any of the analyses and failed to explain differences in social desirability distortion across administration mode. We omitted moderators from all analyses if they did not explain variance in the dependent variable before and after controlling for additional moderators. The following moderators were therefore removed from all analyses: study design, percentage of women in the sample, mean age of the sample, and type of subject population. The lack of relation between these study characteristics and social desirability distortion is informative; differences in research design and sample characteristics do not appear to alter responses across modes of administration. We also removed moderators from the analyses if there was an insufficient number of studies to provide information for a given moderator. In the case of scale reliability, approximately 40% of the sample would have been dropped from the analyses if reliability were included as a moderator. Hence scale reliability was removed from the analyses.

The mean of 581 effect sizes comparing computer and paper-and-pencil instruments was .05 (standard error computed with SUDAAN = .04). These results indicate that computer administration did not have an overall effect on social desirability distortion when compared with paper-and-pencil administration. In contrast, the mean of 92 effect sizes for computerized and face-to-face interviews was −.19 (standard error computed with SUDAAN = .13); this modest effect size suggests that people may make less socially desirable responses in computer instruments than in face-to-face interviews. The computer vs. face-to-face mean effect size (−.19) was significantly less ($p < .05$) than the computer vs. paper-and-pencil mean effect size (.05); this result supports Hypothesis 1, which predicted a near-zero effect size for the computer vs. paper-and-pencil comparison but a substantial effect size for the computer vs. face-to-face comparison in the direction of less social desirability responding on the computer. However, in these comparisons, moderators and dependent variables differed (e.g., comparisons with paper-and-pencil instruments rarely used measures of sensitive personal information whereas comparisons with face-to-face interviews rarely used direct measures of social desirability distortion). Because of these differences, a clear test of Hypothesis 1 could not be made. To understand better what processes may have led to more or less social desirability distortion on the computer, we performed subsequent analyses separately for computer versus paper-and-pencil effect sizes and for computer versus face-to-face interview effect sizes.

Computer Versus Paper-and-Pencil Questionnaires

Table 2 presents the means, standard deviations, and intercorrelations of the variables used to compare computer measures with paper-and-pencil instruments. Unfortunately, SUDAAN does not provide corrected standard errors for Pearson product-moment correlations when there is nonindependence. Although we expected design effects greater than 1 due to correlated observations within study, our analyses with the SUDAAN computer program indicated that nonindependence was not a serious problem for these data (i.e., the design effects in Tables 3 and 4 were not consistently larger than one). Presumably, observed design effects less than unity are due to sampling variability. Because our correlated observations appeared to have little or no effect on the precision of estimated regression coefficients, we used the usual tests for evaluating significance. The significance of the correlations, however, should be evaluated with considerable caution.

Measures of social desirability distortion. If social desirability distortion is affected by computerized administration, this effect should appear in studies using measures designed to assess levels of social desirability distortion. We evaluated, as a group, studies using measures developed to assess social desirability distortion directly (e.g., BIDR) and studies comparing predicted versus actual behavior. All these are computer versus paper-and-pencil comparisons. Table 3 provides a summary of these analyses. Before controlling for moderators, the mean effect size for measures of social desirability distortion was just .01 but after controlling for moderators, the mean effect size was −.39, indicating that measured social desirability distortion was less in the computer than in the paper-and-pencil condition.

Because computer instruments and the interests of researchers changed over time in ways connected with our hypotheses, the year of publication was entered first into the regression equation examining direct measures of social desirability distortion. This variable had a significant effect

Table 2
Descriptive Statistics and Intercorrelations of Study Variables for Computer Versus Paper-and-Pencil Surveys and Questionnaires

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Effect	.05	.52	—																
2. Measure 1	1.57	.49	−.01	—															
3. Measure 2	1.02	.15	−.05	−.18**	—														
4. Measure 3	1.12	.29	.04	−.40**	−.05	—													
5. Measure 4	1.11	.32	.04	−.42**	−.06	−.12**	—												
6. PopD1	1.13	.34	.05	−.02	.17**	.00	−.01	—											
7. PopD2	1.06	.23	−.04	−.01	.16**	−.08*	.01	−.09*	—										
8. PopD3	1.35	.48	−.10*	.14**	−.09*	.05	−.11*	−.29**	−.18**	—									
9. Female	.42	.28	.04	.08	.11*	.05	−.10*	−.19**	−.34**	.05	—								
10. Age	26.64	8.62	−.05	.10*	−.08	.09*	−.05	.40**	.41**	.37**	−.46**	—							
11. Year	86.39	5.52	.05	−.12*	−.03	.29**	−.11*	.03	−.03	.28**	.07	.21**	—						
12. Anony	1.40	.49	.07	−.09	.07	.17*	−.06	.18**	−.19**	.25**	.17*	.08	−.18**	—					
13. Alone	1.78	.49	.02	.05	.28**	−.17*	.09	.12	−.19*	−.29**	−.12	−.41**	−.33**	.01	—				
14. Skip	1.75	.44	−.05	.16**	−.09	.16**	−.08	−.03	−.42**	.33**	.25**	.03	−.25**	.42**	.26**	—			
15. Back	1.54	.50	−.19**	.20**	−.03	−.08	−.02	−.11*	.25**	.64**	−.25**	.76**	.27**	−.24**	−.38**	.13*	—		
16. Design	1.53	.50	−.01	−.10*	.12*	.07	.00	.11*	−.20**	.00	−.21**	−.05	.06	.13*	.48**	.13*	−.11*	—	
17. Reliab	.78	.13	−.18**	.09	.00	.04	−.15**	−.04	−.03	.22**	−.01	.15*	−.04	.03	.13	.24*	.18*	−.07	—

Note. Effect sizes are negative when there was less social desirability distortion on the computer instrument and positive when there was more social desirability distortion on the computer instrument. Variables that were explicitly manipulated (e.g., anonymity, alone, skip, and back) were coded as either 1 or 2. Measure 1 = dummy variable coding for type of instrument, where 2 = personality scale and 1 = all other instrument types; Measure 2 = dummy variable coding for type of instrument, where 2 = behavior inventory and 1 = all other instrument types; Measure 3 = dummy variable coding for type of instrument, where 2 = attitude scale and 1 = all other instrument types; Measure 4 = dummy variable coding for type of instrument, where 2 = nonpsychological, behavioral, and attitudinal scales and 1 = all other instrument types; PopD1 = dummy variable coding for type of population, where 2 = psychiatric population and 1 = all other populations; PopD2 = dummy variable coding for type of population, where 2 = medical population and 1 = all other populations; PopD3 = dummy variable coding for type of population, where 2 = adult population and 1 = all other populations; Female = percentage of women; Age = mean age of participants; Year = year of publication; Anony = anonymity in the computer condition, where 2 = ensured anonymity and 1 = respondent was identified; Alone = respondent alone in the computer condition, where 2 = alone and 1 = not alone; Skip = skip item option on the computer, where 2 = skip option available and 1 = skip option not available; Back = backtracking option on the computer, where 2 = backtracking option available and 1 = backtracking option not available; Design = design of the study, where 2 = between-subjects and 1 = within-subjects; Reliab = reliability of the measure. Because of the clustered nature of the data, the significance of the correlations is approximate.
* $p < .05$. ** $p < .01$.

Table 3
Hierarchical Analyses for Computer Versus Paper-and-Pencil Measures of Social Desirability Distortion

Moderator	R^2	Error *df*	Predicted ES	b^a	*SE* $b^{a,b}$	Design effect[a,b,c]
Step 1	.08	103				
Year of publication				.03**	.01	.32
Recent (1996)			−.08			
Early (1975)			−.74			
Step 2	.28	37				
Anonymity				−.28	.16	.46
Anonymous			−.55			
Identified			−.27			
Alone				−.57**	.19	.46
Alone			−.82			
Not alone			−.25			
Skipping				−.34	.22	1.01
Available			−.54			
Not available			−.20			
Backtracking				−.41**	.15	.42
Available			−.65			
Not available			−.24			

Note. Effect sizes are negative when there was less social desirability distortion on the computer and positive when there was more social desirability distortion on the computer. ES = mean-weighted effect size.
[a] Computed in the final step of the analysis. [b] Computed with SUDAAN. [c] Refers to the extent to which effect sizes within studies were not independent.
* $p < .05$. ** $p < .01$.

(the regression coefficient $b = .03$ with a standard error of .01). To further examine the moderators, we computed the predicted effect sizes for various conditions associated with our hypotheses. To obtain predicted values from the regression equation, we multiplied the mean of each variable by its raw score regression coefficient, except for the variable under consideration. For this variable, we inserted a representative value for each level of the variable. For example, we used 1975 as a date early in the history of research on computer versus paper-and-pencil administration; here the predicted effect size was −.74 whereas the predicted effect size was −.08 for studies published more recently (i.e., 1996). One plausible explanation of this trend is that investigators who published later used better computers and software; they were able to develop computer instruments that more closely matched the format of traditional paper-and-pencil questionnaires, which would be expected to produce more similar results.

After year of study, dummy variables coding for whether or not respondents were told the instrument was anonymous (Hypothesis 2), whether respondents were alone while completing the instrument (Hypothesis 3), and whether or not a skipping and backtracking option was provided on the computer (Hypothesis 4), were entered into the regression equation. This set of variables substantially increased the squared multiple correlation ($\Delta R^2 = .20$). The moderator coding for anonymity was not significant, indicating a lack of support for Hypothesis 2. However, in many studies, subjective anonymity was probably reduced because participants were tested in groups or were observed by an experimenter. The moderator coding for whether respondents were alone was significant ($b = -.57$, with a standard error of .19), indicating less social desirability distortion with computer instruments as compared with the paper-and-pencil instruments when respondents responded alone in both conditions. The predicted effect size for the studies in which respondents completed the instruments alone was −.82 whereas the predicted effect size for the studies in which respondents completed the instruments in the presence of others was −.25. These results support Hypothesis 3; respondents gave less favorable assessments of themselves on the computer when they were alone while completing the instrument.

The moderator coding for computerized backtracking was also significant ($b = -.41$ with a standard error of .15). When computerized versions of paper-and-pencil instruments had more similar formats (e.g., both allowed backtracking), there was less social desirability distortion in the computer condition relative to the paper-and-pencil version (predicted effect size = −.65); however, when backtracking was not available on the computer, this difference was reduced and the responses were more like those in the paper-and-pencil version (predicted effect size = −.24). Note that even when backtracking was not available on the computer, there was less social desirability distortion in the computer condition using these measures of distortion; Hypothesis 4 predicted more social desirability distortion on the computerized assessment in this situation and was consequently not supported.

Reviewers suggested we conduct an additional analysis

comparing the two facets of the BIDR, impression management and self-deception; the literature suggests that impression management is more sensitive to context than self-deception and would show larger effect sizes (Paulhus, 1984). Five studies reported results for both components of the BIDR; one study examined just impression management (Whitener & Klein, 1995), and one examined just self-deception (Finegan & Allen, 1994, Study 3). On the basis of these studies, there does not appear to be a difference in the effect of computerized instruments on impression management as compared with self-deception (impression management $d = -.033$; self-deception $d = .029$, $\chi^2(1, N = 12) = .67$, $p < .42$).

Social desirability distortion inferred from other scales. Table 4 presents a summary of the analysis of studies in which investigators compared social desirability distortion inferred from scores on computer and paper-and-pencil versions of personality scales, behavioral assessments, symptom checklists, and attitude scales. As noted previously, scales were recoded so that higher scores always referred to greater social desirability. Of course, in some of the studies respondents might have had other agendas, such as wanting to fake bad. Therefore, our analysis of these studies must be interpreted with some caution.

Overall, for measures not developed to measure social desirability, the mean effect size for the computer versus paper-and-pencil instrument was .06; with modifiers, the mean effect size for these "indirect" measures was .46 (more distortion in the computer condition). Dummy variables coding for the type of assessment were entered first into the regression equation and found to be nonsignificant ($\Delta R^2 = .01$). That is, there were no significant differences in social desirability distortion when comparing personality, attitude, and behavior scales on computer and paper-and-pencil formats with one another. Year of publication was entered next and was also not significant. In the last step, we entered dummy variables coding for anonymity (Hypothesis 2), being alone (Hypothesis 3), and ability to skip and backtrack (Hypothesis 4). This set of variables increased the squared multiple correlation ($\Delta R^2 = .07$).

In general, these studies suggest there might be more social desirability distortion on the computer than on the paper-and-pencil instruments. However, the relative size of the effects varied in accord with the predictions of Hypoth-

Table 4
Hierarchical Analyses of Social Desirability Distortion in Computer Versus Paper-and-Pencil Personality, Behavior, and Attitude Measures

Moderator	R^2	Error *df*	Predicted ES	b[a]	*SE* b[a,b]	Design effect[a,b,c]
Step 1	.01	472				
Measure 1[d]				.10	.16	.25
Personality scale			.49			
All other measures			.39			
Measure 2				.41	.44	1.78
Behavior inventory			.86			
All other measures			.45			
Measure 3				−.12	.25	.19
Attitude scale			.36			
All other measures			.48			
Step 2	.01	471				
Year of publication				.01	.01	.22
Recent (1996)			.57			
Early (1975)			.35			
Step 3	.08	88				
Anonymity				−.37*	.15	.06
Anonymous			.25			
Identified			.62			
Alone				−.53**	.17	.18
Alone			.12			
Not alone			.65			
Skipping				.13	.11	.06
Available			.49			
Not available			.36			
Backtracking				−.71**	.12	.05
Available			.16			
Not available			.87			

Note. Effect sizes are negative when there was less social desirability distortion on the computer and positive when there was more social desirability distortion on the computer. ES = mean-weighted effect size.
[a] Computed in the final step of the analysis. [b] Computed with SUDAAN. [c] Refers to the extent to which effect sizes within studies were not independent. [d] Dummy variables coding for the type of measure.
* $p < .05$. ** $p < .01$.

eses 2, 3, and 4. The anonymity moderator was significant ($b = -.37$ with a standard error of .15), indicating that when respondents were assured of anonymity they showed somewhat more social desirability distortion on the computer (predicted effect size = .25), but when identified, they showed much more social desirability distortion on the computer than in the paper-and-pencil condition (predicted effect size = .62). Also, the moderator coding for whether respondents were alone was significant ($b = -.53$ with a standard error of .17); when respondents completed the assessments alone they showed only a little more social desirability distortion on the computer (predicted effect size = .12), but when they were tested in the presence of others, they showed much more social desirability distortion on the computer than on the paper-and-pencil instrument (predicted effect size = .65). Finally, being able to backtrack on the computer significantly affected the degree of social desirability distortion on these measures ($b = -.71$ with a standard error of .12). When backtracking on the computer was allowed, there was a little more social desirability distortion on the computer (predicted effect size = .16) but when backtracking was not allowed, there was much more social desirability distortion on the computer (predicted effect size = .87).

We were unable to test Hypothesis 5, that asking for sensitive personal information reduces comparative social desirability distortion in computer instruments versus paper-and-pencil questionnaires. We coded only 20 of the 456 comparisons between paper-and-pencil and computer instruments as using measures of sensitive personal information such as illegal behavior, criminal history, or sexual practices. Those few studies that used sensitive questions also tended to be among the earlier studies ($r = -.31$).

Computer Versus Face-to-Face Interviews

In Table 5, we present means, standard deviations, and intercorrelations of the variables used to compare computer-based interviews with face-to-face interviews. As stated previously, the significance of these correlations should be evaluated with caution due to the clustered nature of the data.

In Table 6, we present a summary of the analyses comparing computer assessments with face-to-face interviews. In contrast with the analyses described in Tables 3 and 4, the design effects shown in Table 6 are consistently larger than one. Evidently the procedures and instruments used in the interview studies created stronger correlations among conditions and measures than they did in studies comparing computers with paper-and-pencil surveys.

Dummy variables coding for the type of assessment were entered first into the regression and accounted for 15% of the variance in the dependent variable. Next, the year of publication was entered into the equation; the increase in variance explained was minimal although significant ($\Delta R^2 = .04$). Finally, dummy variables coding for anonymity, alone, skipping, and backtracking were entered into the equation; a significant and substantial increase in the squared multiple correlation was observed ($\Delta R^2 = .24$).

One of the dummy variables coding for type of assessment was significant ($b = 1.12$ with a standard error of .46). That is, respondents displayed relatively less social desirability distortion on the computer when the measure was a behavioral measure, symptom checklist, or an attitude scale (predicted effect size = −.51) and more social desirability distortion in the computer interview when the measure was a personality scale (predicted effect size = .73). This finding is relevant to Hypothesis 5, that social desirability distortion would be reduced on the computer when the instrument requested highly sensitive personal information. Many of the behavioral measures and symptom checklists used in comparisons of face-to-face interviews with computer interviews asked for somewhat or highly sensitive personal information such as information about the respondent's medical status, mental health, or criminal record, as well as information regarding illegal drug use and risky sexual behavior. In our data set, the dummy variable coding for highly sensitive information was highly correlated with the dummy variable coding for behavioral measures and symptom checklists ($r = .92$, $p < .01$). The finding that (a) less social desirability distortion occurred in the computer interview when the measure was a behavioral measure, symptom checklist, or attitude scale coupled with (b) the near perfect correlation between the sensitivity of the measure and the use of these instruments suggest that participants may have been less concerned with social desirability when they responded to sensitive items on the computer (Hypothesis 5).

In the analyses presented in Table 6, year of publication was also significant ($b = .13$ with a standard error of .05). In studies published recently, respondents engaged in relatively more social desirability distortion on the computer than in the face-to-face interview (predicted effect size = .79 for a study published in 1995); in studies published in previous years, there was much less social desirability distortion on the computer than in the face-to-face interview (predicted effect size = −1.03 for a study published in 1981). However, earlier studies were more likely to include measures of highly sensitive information and to have an interface that differed from traditional measures. Adding the final set of variables in the third step of the analysis resulted in a large increase in R^2 ($\Delta R^2 = .24$), but due to missing data only 37 degrees of freedom remained.

Discussion

Many investigators have speculated that particular attributes of computer instruments, such as the display of

Table 5
Descriptive Statistics and Intercorrelations of Study Variables for Computer Versus Face-to-Face Interviews

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Effect	−.19	.78	—															
2. Measure 1	1.34	.49	.05	—														
3. Measure 2	1.32	.47	−.35**	−.48**	—													
4. Measure 3	1.28	.45	.29**	−.45**	−.43**	—												
5. PopD1	1.45	.50	−.32**	.42**	.24*	−.56**	—											
6. PopD2	1.03	.18	.00	−.13	.27**	−.12	−.16	—										
7. PopD3	1.21	.41	.07	−.14	.00	.16	−.46**	−.09	—									
8. Female	.56	.26	.16	.44**	−.32**	−.09	.55**	−.30*	−.33**	—								
9. Age	25.81	8.38	−.31**	.12	.32**	−.40**	.30**	.18	.31**	−.53**	—							
10. Year	88.09	4.87	.31**	.17	−.26*	.26*	−.07	−.16	.08	.08	.15	—						
11. Anony	1.41	.50	.41**	−.10	−.29*	.48**	−.21	−.18	−.12	.59**	−.73**	.07	—					
12. Alone	1.91	.38	.01	−.06	−.25*	.25*	.12	.05	−.57**	.09	−.32**	−.16	.31*	—				
13. Skip	1.44	.50	.34**	.10	−.19	.08	.03	.14	−.24	.57**	−.60**	.19	.89**	.30*	—			
14. Back	1.38	.49	−.11	.54**	−.32*	−.34*	−.01	—	.21	.21	.54**	.28*	−.71**	−.44**	−.51**	—		
15. Design	1.37	.49	.31**	−.12	.06	−.03	−.05	.01	−.17	.27*	−.58**	.07	.76**	.38**	.69**	−.54**	—	
16. Reliab	.82	.08	−.31	.53**	−.16	−.42**	.06	.12	.40*	−.70**	.73**	.54**	−.82**	−.47**	−.43**	.85**	−.43**	—

Note. Effect sizes are negative when there was less social desirability distortion on the computer instrument and positive when there was more social desirability distortion on the computer instrument. Variables that were explicitly manipulated (e.g., anonymity, alone, skip, and back) were coded as either a 1 or a 2. Measure 1 = dummy variable coding for type of instrument, where 2 = personality scale and 1 = all other instrument types; Measure 2 = dummy variable coding for type of instrument, where 2 = behavior inventory and 1 = all other instrument types; Measure 3 = dummy variable coding for type of instrument, where 2 = attitude scale and 1 = all other instrument types; PopD1 = dummy variable coding for type of population, where 2 = psychiatric population and 1 = all other populations; PopD2 = dummy variable coding for type of population, where 2 = medical population and 1 = all other populations; PopD3 = dummy variable coding for type of population, where 2 = adult population and 1 = all other populations; Female = percentage of women; Age = mean age of participants; Year = year of publication; Anony = anonymity in the computer condition, where 2 = ensured anonymity and 1 = respondent was identified; Alone = respondent alone in the computer condition, where 2 = alone and 1 = not alone; Skip = skip item option on the computer, where 2 = skip option available and 1 = skip option not available; Back = backtracking option on the computer, where 2 = backtracking option available and 1 = backtracking option not available; Design = design of the study, where 2 = between-subjects and 1 = within-subjects; Reliab = reliability of the measure. Because of the clustered nature of the data, the significance of the correlations is approximate.
* $p < .05$. ** $p < .01$.

Table 6
Hierarchical Analyses of Social Desirability Distortion in Computer Versus Face-to-Face Interviews

Moderator	R^2	Error *df*	Predicted ES	b^a	*SE* $b^{a,b}$	Design effect[a,b,c]
Step 1	.15	88				
Measure 1[d]				1.12*	.46	1.05
Personality scale			.73			
All other measures			−.51			
Measure 2				1.24	.65	1.99
Behavior inventory			.63			
All other measures			−.49			
Measure 3				.31	.54	1.21
Attitude scale			.11			
All other measures			−.20			
Step 2	.19	87				
Year of publication				.13*	.05	1.11
Recent (1996)			.79			
Early (1975)			−1.03			
Step 3	.43	37				
Anonymity				.74	.48	1.37
Anonymous			.33			
Identified			−.41			
Alone				.46	.29	1.72
Alone			−.03			
Not alone			−.49			
Skipping				−.26	.29	1.99
Available			−.26			
Not available			.00			
Backtracking				−.14	.44	.83
Available			−.20			
Not available			−.06			

Note. Effect sizes are negative when there was less social desirability distortion on the computer and positive when there was more social desirability distortion on the computer. ES = mean-weighted effect size. [a] Computed in the final step of the analysis. [b] Computed with SUDAAN. [c] Refers to the extent to which effect sizes within studies were not independent. [d] Dummy variables coding for the type of measure.
* $p < .05$. ** $p < .01$.

plain text on a screen, the ephemeral nature of responses, the absence of social context cues, and constraints on how respondents view and answer items, can change respondents' perceptions of the computer test situation and lead to important differences for the computer instrument. However, considerable debate has surrounded which attributes have real importance for social desirability distortion. Our analysis of the literature comparing social desirability distortion in computer and traditional noncognitive instruments over the last 30 years suggests why so many investigators have described this literature as "mixed" or "conflicting" (e.g., Rosenfeld et al., 1996). We found that using a computer instrument per se has no consistent effect on distortion; the effect, if any, depends on what the instrument measures and on moderating factors such as whether respondents are tested alone or in the presence of others. However, the number of studies has reached sufficient size to sort this literature along practical and theoretical dimensions and to point out where research is needed.

The evidence suggests that for practical decision making in many testing situations, computer and paper-and-pencil scales can be expected to give similar mean results. However, there are other facets to measurement equivalence that we did not examine in this meta-analysis. Moreover, studies in which social desirability distortion was measured directly suggest that when computer instruments are administered to respondents who are alone and when respondents can respond fairly freely (e.g., backtrack to previous responses), they may feel particularly comfortable or less wary in giving socially undesirable answers (predicted effect size = −.39 when means for all moderators are inserted into the regression equation). Assurances of anonymity on direct measures did not significantly alter respondents' distortion on the computer, but anonymity instructions may have been weakened in many of these studies because respondents answered in the presence of an experimenter or other respondents. We were unable to test the interaction of anonymity instructions and being alone; doing so would have entailed a loss of 80% of the effect sizes due to missing data.

Personality tests such as the MMPI, employment tests, and other standardized scales are developed to minimize or to control for social desirability distortion, but they often include many negative items (I have trouble sleeping; I feel lonely; I have imagined killing someone) that require people

to deny negative attributes in order to appear normal or socially desirable (Walsh, 1990). Computer instruments that substituted for these scales tended to increase distortion on the computer (predicted effect size = .46 when means for all moderators are inserted into the regression equation) but the effect was small when respondents were anonymous or alone, and when they could backtrack to previous responses. By contrast, these computer instruments dramatically increased respondents' unwillingness to reveal personal weaknesses—heightened social desirability distortion—when respondents were identified or in the presence of others, or could not backtrack to previous responses. One possible explanation of this finding is that respondents had expectations about the evaluative purposes of many of these instruments (such as the MMPI), raising their concern about how their data could be used if they were kept on a computer. Consistent with this possibility, Rosenfeld et al.'s (1996) study suggests that instructions reminding vulnerable respondents that their data will be kept on a computer increases distortion relative to that given with a paper-and-pencil scale.

A key finding of our analysis was that computer instruments reduced social desirability distortion when these instruments were used as a substitute for face-to-face interviews, particularly when the interviews were asking respondents to reveal highly sensitive personal behavior, such as whether they used illegal drugs or engaged in risky sexual practices. The most obvious interpretation of this finding is that computer interviews reduced social desirability distortion as compared with face-to-face interviews for much the same reason paper-and-pencil questionnaires do—they are self-administered and more removed from the observation of an interviewer and from social cues that arouse evaluation apprehension and reduce neutrality (Sudman & Bradburn, 1974). In this regard, it may seem curious that face-to-face interviews are so frequently used to gather highly sensitive information. The answer seems to be that face-to-face interviews are motivating, reduce nonresponse, and encourage longer, more elaborated answers; in effect, distortion is traded for more complete responses (Sudman & Bradburn, 1974). Further, distortion can be reduced with assurances of confidentiality (Woods & McNamara, 1980). Computer interviews potentially change this equation by reducing distortion as paper-and-pencil instruments do while increasing response completeness over paper-and-pencil instruments. Some investigators who have examined respondent satisfaction have reported that most respondents enjoy completing the computer instruments (Binik, Meana, & Sand, 1994) and they provide longer, more elaborated answers on the computer (Sproull & Kiesler, 1986).

The SUDAAN computer program (Shah et al., 1995) allowed us to use all the data that we were able to obtain. Multiple (and nonindependent) effect sizes were very common; in fact, the 61 studies contained 673 effect sizes, or about 11 effect sizes per study. The amount of information available for analysis would have been greatly reduced if we had attempted to restrict our sample to statistically independent cases. In some cases, the SUDAAN analyses revealed that such a restriction was unnecessary; few of the paper-and-pencil versus computer design effects greatly exceeded 1.0 but the face-to-face interview versus computer comparisons told a different story. Without the use of SUDAAN and its computation of design effects, we would not have known which multiple effect sizes reported in this literature were independent and which were not.

As a check on our use of the SUDAAN program, we performed a parallel set of regression analyses using SPSS (SPSS, Inc., 1996). As explained by Shah et al. (1995), these analyses should result in identical estimates of regression coefficients, and they did. The standard errors of the regression coefficients differed to some extent, however, because SUDAAN does not assume observations are independent. Corresponding to the findings about design effects, the SPSS standard errors were generally similar in size to the SUDAAN standard errors. In summary, we encourage meta-analytic researchers to analyze all the data available to them using statistical procedures that make appropriate assumptions about nonindependent observations.

Limitations

Our conclusions should be interpreted with caution. In some cases, one study characteristic (e.g., highly sensitive information) was highly correlated with another study characteristic (e.g., use of symptom checklist format). Given such correlations, it is impossible to untangle causal relations and we suggest that future research should address such confounds by using appropriate experimental designs. Missing data were also a considerable problem for our analyses. It was particularly difficult to get complete information on the anonymity, alone, skipping, and backtracking variables. We attempted to contact authors in many cases, but despite our best efforts much information remained unavailable. Furthermore, as stated previously, approximately 40% of the sample did not report reliability estimates. We decided not to drop studies when the reliability of measures was omitted, but that meant we could not assess whether reliability was an artifact in the analysis.

Directions for Research

The past 30 years of research on computerized noncognitive instruments generally has had the somewhat modest goal of evaluating the similarity of computer instruments to traditional instruments. One problem in much of this research is the implicit assumption that there is a simple relationship between threat and distortion or safety and accuracy. On the contrary, social desirability distortion is

probably nonlinearly related to many descriptions of the self. That is, reporting too much of any good trait looks like immodesty and reporting too much of any bad trait looks like malingering. Hence someone giving a strategically socially desirable response could actually report less of a good trait than someone less strategically but fancifully giving an aggrandizing self-description. It is possible that respondents may be more prone to brag, to fantasize, or to malinger when they know their answers cannot be checked (e.g., on an anonymous Web survey). It is also possible that, in similar circumstances, respondents may attain a feeling of well-being that in turn increases overconfidence or positive judgments. To our knowledge, no researchers have investigated nonlinear social desirability effects of computer and traditional instruments. New research on computer instruments might well be aimed at a better understanding of respondents' motivations to distort or to give accurate answers. Research is also needed on cognitive biases that could influence distortion. This research would aid in our understanding of patterns and implications of distortion in computer and traditional instruments, and to our more general understanding of respondents' reactions to assessment.

A related problem in the literature is that few investigators examined distortion in computer and traditional instruments when the assessment really mattered. The literature on social desirability distortion suggests that respondents are likely to distort their responses in a favorable direction when the instrument matters, as in a job interview (Douglas et al., 1996; Kluger & Colella, 1993). Other possibilities are that respondents will fake bad when they want help or attention; that they will respond accurately when they are rewarded for accuracy; that they will respond playfully when they are encouraged to have fun. Yet virtually all of the research comparing computer and paper-and-pencil scales was conducted in settings where there were no clear personal consequences for the respondents of completing the measures. At the time this meta-analysis was conducted, no published research directly tested the effect of how respondents' data would be used on their social desirability distortion in computer instruments. Wilkerson, Nagao, and Martin (1997) recently used a role playing exercise to manipulate the importance of the instrument; they report that their "job screening" scale elicited higher social desirability distortion than their "consumer survey" and that mode of administration had no differential effect. However, we were unable to locate any studies comparing computerized and paper-and-pencil instruments when real personal consequences of a test were measured or manipulated.

A third problem in the literature is that investigators have rarely thoroughly assessed respondents' perceptions of anonymity, privacy or confidentiality, and neutrality or tried to link these perceptions with particular attributes of the context or interface. The possibility that computerization of a questionnaire or interview could lead to a (possibly illusory) feeling of anonymity, privacy, or neutrality, and therefore could encourage more honest reporting of sensitive information, has stimulated the development of computer instruments to collect highly sensitive information such as risk behaviors of blood donors (AIR, 1993; Locke et al., 1992). Our data suggest that, at least in the domain of interviews to collect sensitive or intimate information, this is a promising direction. However, little research has been conducted on which aspects of the computer interface or computer test context increase or decrease perceptions of anonymity, privacy, or neutrality. Most investigators only asked respondents if they were comfortable with the computer (or traditional) instrument. Honaker (1988), in a review of the computerized MMPI literature, argued that the computer interface can have important effects on respondents and he urged researchers comparing computerized and traditional instruments to provide details of the interface so that, in the future, researchers could identify the computer-user interaction features that lead to instrument equivalence or nonequivalence.

The literature of the last 30 years has quieted concern about the appropriateness of noncognitive computer instruments, but research on social desirability distortion in computer assessment will continue as possibilities for new kinds of computer assessment open. Speech simulation and speech understanding technology have reached the point that real time or automated audio and video interviews (Alemi & Higley, 1995; Johnston & Walton, 1995; Tourangeau & Smith, 1996; Turner et al., 1998) and interviews by digitized characters (e.g., Sproull, Subramani, & Kiesler, 1996) can be used in place of traditional face-to-face interviews. This technology allows for computer "interviews" with self-administration, anonymity, and no presence of others, as well as contingent questioning (e.g., branching). Such computer-based interviews could be compared with traditional interviews or with paper questionnaires, avoiding some of the confounds (e.g., self-administration) in earlier studies. Early evidence suggests that audio computer-assisted self-interviewing with anonymity instructions may reduce social desirability distortion of sensitive information over the distortion from the types of computer instruments evaluated in this article as well as from face-to-face interviews and paper-and-pencil instruments (Tourangeau & Smith, 1996; Turner et al., 1998). However, the reliability of this effect and mechanisms responsible for it await further study.

References

Studies preceded by an asterisk were included in the meta-analysis.

*Alemi, F., & Higley, P. (1995). Reaction to "talking" computers assessing health risks. *Medical Care, 33,* 227–233.

American Institutes for Research. (1993, July 30). *Increasing the*

safety of the blood supply by screening donors more effectively (Contract No. 223-91-1002). Washington, DC: Author.

*Angle, H. V., Johnsen, T., Grebenkemper, N. S., & Ellinwood, E. H. (1979). Computer interview support for clinicians. *Professional Psychology, 10,* 49–57.

*Barron, M. R., Daniels, J. L., & O'Toole, W. M. (1987). The effect of computer-conducted versus counselor-conducted initial intake interviews on client expectancy. *Computers in Human Behavior, 3,* 21–28.

Binik, Y. M., Meana, M., & Sand, N. (1994). Interaction with a sex-expert system changes attitudes and may modify sexual behavior. *Computers in Human Behavior, 10,* 395–410.

*Biskin, B. H., & Kolotkin, R. L. (1977). Effects of computerized administration on scores on the Minnesota Multiphasic Personality Inventory. *Applied Psychological Measurement, 1,* 543–549.

*Booth-Kewley, S., Edwards, J. E., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology, 77,* 562–566.

Cannavale, F. J., Scarr, H. A., & Pepitone, A. (1970). Deindividuation in the small group: Further evidence. *Journal of Personality and Social Psychology, 16,* 141–147.

*Canoune, H. L., & Leyhe, E. W. (1985). Human versus computer interviewing. *Journal of Personality Assessment, 49,* 103–106.

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

*Carr, A. C., Ghosh, A., & Ancill, R. J. (1983). Can a computer take a psychiatric history? *Psychological Medicine, 13,* 151–158.

Catania, J. A., Gibson, D. R., Chitwood, D. D., & Coates, T. J. (1990). Methodological problems in AIDS behavioral research: Influences on measurement error and participation bias in studies of sexual behavior. *Psychological Bulletin, 108,* 339–362.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Crowne, D., & Marlowe, D. (1964). *The approval motive.* New York: Wiley.

*Davis, C., & Cowles, M. (1989). Automated psychological testing: Method of administration, need for approval, and measures of anxiety. *Educational and Psychological Measurement, 49,* 311–320.

*Davis, L. J., Jr., Hoffmann, N. G., Morse, R. M., & Luehr, J. G. (1992). Substance Use Disorder Diagnostic Schedule (SUDDS): The equivalence and validity of a computer-administered and an interviewer-administered format. *Alcoholism: Clinical and Experimental Research, 16,* 250–254.

Diener, E., Fraser, S., Beaman, A. L., & Kelem, R. T. (1976). Effects of deindividuating variables on stealing by Halloween trick-or-treaters. *Journal of Personality and Social Psychology, 33,* 178–183.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996, August). The validity of non-cognitive measures decays when applicants fake. *Proceedings of the 56th Annual Meeting of the Academy of Management* (pp. 127–131). Cincinnati, OH.

Dubrovsky, V., Kiesler, S., & Sethna, B. (1991). The equalization phenomenon: Status effects in computer-mediated and face-to-face decision making groups. *Human Computer Interaction, 6,* 119–146.

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology, 58,* 568–581.

Edwards, A. L. (1957). *The social desirability variable in personality research and assessment.* New York: Dryden Press.

*Erdman, H., Klein, M. H., & Greist, J. H. (1983). The reliability of a computer interview for drug use/abuse information. *Behavior Research Methods and Instrumentation, 15,* 66–68.

*Erdman, H. P., Klein, M. H., Greist, J. H., Skare, S. S., Husted, J. J., Robins, L. N., Helzer, J. E., Goldring, E., Hamburger, M., & Miller, P. (1992). A comparison of two computer-administered versions of the NIMH Diagnostic Interview Schedule. *Journal of Psychiatric Research, 26,* 85–95.

*Evan, W. M., & Miller, J. R., III (1969). Differential effects on response bias of computer vs. conventional administration of a social science questionnaire: An exploratory methodological experiment. *Behavioral Science, 14,* 216–227.

*Farrell, A. D., Camplair, P. S., & McCullough, L. (1987). Identification of target complaints by computer interview: Evaluation of the computerized assessment system for psychotherapy evaluation and research. *Journal of Consulting and Clinical Psychology, 55,* 691–700.

*Ferriter, M. (1993). Computer aided interviewing in psychiatric social work. *Computers in Human Services, 91,* 59–66.

*Finegan, J. E., & Allen, N. J. (1994). Computerized and written questionnaires: Are they equivalent? *Computers in Human Behavior, 10,* 483–496.

*French, C. C., & Beaumont, J. G. (1989). A computerized form of the Eysenck Personality Questionnaire: A clinical study. *Personality and Individual Differences, 10,* 1027–1032.

*George, C. E., Lankford, J. S., & Wilson, S. E. (1992). The effects of computerized versus paper-and-pencil administration on measures of negative affect. *Computers in Human Behavior, 8,* 203–209.

Greist, J. H., & Klein, M. K. (1981). Computers in psychiatry. In S. Arieti & H. K. H. Brodie (Eds.), *American handbook of psychiatry: Vol. 7* (2nd ed., pp. 750–777). New York: Basic Books.

*Greist, J. H., Klein, M. H., Erdman, H. P., Bires, J. K., Bass, S. M., Machtinger, P. E., & Kresge, D. G. (1987). Comparison of computer- and interviewer-administered versions of the Diagnostic Interview Schedule. *Hospital and Community Psychiatry, 38,* 1304–1311.

*Greist, J. H., Klein, M. H., Van Cura, L. J., & Erdman, H. P. (1975). Computer interview questionnaires for drug use/abuse. In D. J. Lettieri (Ed.), *Predicting drug abuse: A review of issues, methods, and correlates.* Rockville, MD: National Institute of Drug Abuse.

Griffith, T. L., & Northcraft, G. B. (1994). Distinguishing between the forest and the trees: Media, features, and methodology in electronic communication research. *Organization Science, 5,* 272–285.

*Hart, R. R., & Goldstein, M. A. (1985). Computer-assisted psychological assessment. *Computers in Human Services, 1,* 69–75.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

*Hinkle, J. S., Sampson, J. P., & Radonsky, V. (1991). Computer-assisted versus traditional and paper assessment of personal problems in a clinical population. *Computers in Human Behavior, 7,* 237–242.

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

*Honaker, L. M., Harrell, T. H., & Buffaloe, J. D. (1988). Equivalency of microtest computer MMPI administration for standard and special scales. *Computers in Human Behavior, 4,* 323–337.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Johnston, J., & Walton, C. (1995). Reducing response effects for sensitive questions: A computer-assisted self interview with audio. *Social Science Computer Review, 13,* 305–319.

*Kantor, J. (1991). The effects of computer administration and identification on the Job Descriptive Index (JDI). *Journal of Business and Psychology, 5,* 309–323.

*Katz, L., & Dalby, J. T. (1981). Computer and manual administration of the Eysenck Personality Inventory. *Journal of Clinical Psychology, 37,* 586–592.

*Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly, 50,* 402–413.

Kiesler, S., & Sproull, L. S. (1992). Group decision making and communication technology. *Organizational Behavior and Human Decision Processes, 52,* 96–123.

Kiesler, S., Walsh, J., & Sproull, L. (1992). Computer networks in field research. In F. B. Bryant, J. Edwards, S. Tindale, E. Posavac, L. Heath, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Methodological issues in applied social research* (pp. 239–268). New York: Plenum.

*King, W. C., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80,* 643–651.

Kluger, A., & Colella, A. (1993). Beyond the mean bias: The effect of warning against faking on biodata item variances. *Personnel Psychology, 46,* 763–780.

*Kobak, K. A., Reynolds, W. M., & Greist, J. H. (1993). Development and validation of a computer-administered version of the Hamilton Anxiety Scale. *Psychological Assessment, 5,* 487–492.

*Kobak, K. A., Reynolds, W. M., Rosenfeld, R., & Greist, J. H. (1990). Development and validation of a computer-administered version of the Hamilton Depression Scale. *Psychological Assessment, 2,* 56–63.

*Koson, D., Kitchen, M., Kochen, M., & Stodolsky, D. (1970). Psychological testing by computer: Effect on response bias. *Educational and Psychological Measurement, 30,* 803–810.

*Lambert, M. E., Andrews, R. H., Rylee, K. R., & Skinner, J. R. (1987). Equivalence of computerized and traditional MMPI administration with substance abusers. *Computers in Human Behavior, 3,* 139–143.

*Lankford, J. S., Bell, R. W., & Elias, J. W. (1994). Computerized versus standard personality measures: Equivalency, computer anxiety, and gender differences. *Computers in Human Behavior, 10,* 497–510.

*Lautenschlager, G. J., & Flaherty, V. L. (1990). Computer administration of questions: More desirable or more social desirability? *Journal of Applied Psychology, 75,* 310–314.

*Levine, S., Ancill, R. J., & Roberts, A. P. (1989). Assessment of suicide risk by computer-delivered self-rating questionnaire: Preliminary findings. *Acta Psychiatrica Scandinavica, 80,* 216–220.

*Liefeld, J. P. (1988). Response effects in computer-administered questioning. *Journal of Marketing Research, 25,* 405–409.

Lindskold, S., & Finch, M. L. (1982). Anonymity and the resolution of conflicting pressures from the experimenter and from peers. *Journal of Psychology, 112,* 79–86.

*Locke, S. D., & Gilbert, B. O. (1995). Method of psychological assessment, self-disclosure, and experiential differences: A study of computer, questionnaire, and interview assessment formats. *Journal of Social Behavior and Personality, 10,* 255–263.

*Locke, S. E., Kowaloff, H. B., Hoff, R. G., Safran, C., Popovsky, M. A., Cotton, D. J., Finkelstein, D. M., Page, P. L., & Slack, W. V. (1992). Computer-based interview for screening blood donors for risk of HIV transmission. *Journal of the American Medical Association, 268,* 1301–1305.

*Lucas, R. W., Mullin, P. J., Luna, C. B. X., & McInroy, D. C. (1977). Psychiatrists and a computer as interrogators of patients with alcohol-related illnesses: A comparison. *British Journal of Psychiatry, 131,* 160–167.

*Lukin, M. E., Dowd, E. T., Plake, B. S., & Kraft, R. G. (1985). Comparing computerized versus traditional psychological assessment. *Computers in Human Behavior, 1,* 49–58.

*Lushene, R. E., O'Neil, H. F., Jr., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment, 38,* 353–361.

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

Matarazzo, J. M. (1983). Computerized psychological testing. *Science, 221,* p. 323.

Mathes, E. W., & Guest, T. A. (1976). Anonymity and group antisocial behavior. *Journal of Social Psychology, 100,* 257–262.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

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

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81,* 660–679.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46,* 598–609.

*Potosky, D., & Bobko, P. (1997). Computer versus paper-and-pencil administration mode and response distortion in noncognitive selection tests. *Journal of Applied Psychology, 82,* 293–299.

*Rezmovic, V. (1977). The effects of computerizing experimentation on response bias variance. *Research Methods and Instrumentation, 9,* 144–147.

*Ridgway, J., MacCulloch, M. J., & Mills, H. E. (1982). Some experiences in administering a psychometric test with a light pen and microcomputer. *International Journal of Man-Machine Studies, 17,* 265–278.

*Robinson, R., & West, R. (1992). A comparison of computer and questionnaire methods of history-taking in a genito-urinary clinic. *Psychology and Health, 6,* 77–84.

Robinson, T. N., & Walters, P. A. (1986). Health-Net: An interactive computer network for campus health promotion. *Journal of American College Health, 34,* 284–285.

*Rosenfeld, P., Booth-Kewley, S., Edwards, J. E., & Thomas, M. D. (1996). Responses on computer surveys: Impression management, social desirability, and the Big Brother syndrome. *Computers in Human Behavior, 12,* 263–274.

*Rosenfeld, P., Doherty, L. M., Vicino, S. M., Kantor, J., & Greaves, J. (1989). Attitude assessment in organizations: Testing three microcomputer-based survey systems. *The Journal of General Psychology, 116,* 145–154.

*Rosenfeld, P., Giacalone, R. A., Knouse, S. B., Doherty, L. M., Vicino, S. M., Kantor, J., & Greaves, J. (1991). Impression management, candor, and microcomputer-based organizational surveys: An individual differences approach. *Computers in Human Behavior, 7,* 23–32.

Rosenfeld, P., Giacalone, R. A., & Riordan, C. A. (1992). *Impression management in organizations: Theory, measurement, practice.* London: Rowledge.

*Rosenfeld, R., Dar, R., Anderson, D., Kobak, K. A., & Greist, J. H. (1992). A computer-administered version of the Yale–Brown Obsessive–Compulsive Scale. *Psychological Assessment, 4,* 329–332.

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

*Schuldberg, D. (1988). The MMPI is less sensitive to the automated testing format than it is to repeated testing: Item and scale effects. *Computers in Human Behavior, 4,* 285–298.

*Scissons, E. H. (1976). Computer administration of the California Psychological Inventory. *Measurement and Evaluation in Guidance, 9,* 22–25.

Shah, B. V., Barnwell, B. G., & Bieler, G. S. (1995). *SUDAAN User's Manual: Software for analysis of correlated data.* Research Triangle Park, NC: Research Triangle Institute.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications.* London: Wiley.

Singer, J., Brush, C., & Lublin, S. (1965). Some aspects of deindividuation: Identification and conformity. *Journal of Experimental Social Psychology, 1,* 356–378.

*Skinner, H. A., & Allen, B. A. (1983). Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug, and tobacco use. *Journal of Consulting and Clinical Psychology, 51,* 267–275.

Slack, W. V., Hicks, G. P., Reed, C. E., & Van Cura, L. J. (1966). A computer-based medical history system. *New England Journal of Medicine, 274,* 194–198.

Smith, P. C., Kendall, L., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago: Rand McNally.

Smith, R. E. (1963). Examination by computer. *Behavioral Science, 8,* 76–79.

Snell, A. F., & McDaniel, M. A. (1998, April). *Faking: Getting data to answer the right questions.* Paper presented at the 13th Annual Conference for the Society for Industrial and Organizational Psychology, Dallas, TX.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26,* 261–271.

*Sproull, L. (1986). Using electronic mail for data collection in organizational research. *Academy of Management Journal, 29,* 159–169.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32,* 1492–1512.

Sproull, L., & Kiesler, S. (1991). *Connections: New ways of working in the networked organization.* Cambridge, MA: MIT Press.

Sproull, L., Subramani, M., & Kiesler, S. (1996). When the interface is a face. *Human–Computer Interaction, 11,* 97–124.

SPSS, Inc. (1996). *SPSS for Windows, Rel. 7.0.* Chicago: Author.

Sudman, S., & Bradburn, N. N. (1974). *Response effects in surveys.* Chicago: Aldine.

Synodinos, N. E., & Brennan, J. M. (1988). Computer interactive interviewing in survey research. *Psychology and Marketing, 5,* 117–137.

*Synodinos, N. E., Papacostas, C. S., & Okimoto, G. M. (1994). Computer administered versus paper-and-pencil surveys and the effect of sample selection. *Behavior Research Methods, 26,* 395–401.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format and question context. *Public Opinion Quarterly, 60,* 275–304.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998, May). Adolescent sexual behavior, drug use, & violence: Increased reporting with computer survey technology. *Science, 280,* 867–873.

*Vansickle, T. R., Kimmel, C., & Kapes, J. T. (1989). Test–retest equivalency of the computer-based and paper-and-pencil versions of the Strong–Campbell Interest Inventory. *Measurement and Evaluation in Counseling and Development, 22,* 88–93.

Walsh, J. A. (1990). Comment on social desirability. *American Psychologist, 45,* 289–290.

*Waterton, J. J., & Duffy, J. C. (1984). A comparison of computer interviewing techniques and traditional methods in the collection of self-report alcohol consumption data in a field survey. *International Statistical Review, 52,* 173–182.

*Watson, C. G., Manifold, V., Klett, W. G., Brown, J., Thomas, D., & Anderson, D. (1990). Comparability of computer- and booklet-administered Minnesota Multiphasic Personality Inventories among primarily chemically dependent patients. *Psychological Assessment, 2,* 276–280.

*White, D. M., Clements, C. B., & Fowler, R. D. (1985). A comparison of computer administration with standard adminis-

tration of the MMPI. *Computers in Human Behavior, 1,* 153–162.

White, M. J. (1977). Counternormative behavior as influenced by deindividuating conditions and reference group salience. *Journal of Personality and Social Psychology, 103,* 73–90.

*Whitener, E. M., & Klein, H. J. (1995). Equivalence of computerized and traditional research methods: The roles of scanning, social environment, and social desirability. *Computers in Human Behavior, 11,* 65–75.

Wilkerson, J. M., Nagao, D. H., & Martin, C. L. (1997, August). *Socially desirable responding in computerized questionnaires: When context matters more than the medium.* Paper presented at the Academy of Management Annual Meeting, Boston.

*Wilson, F. R., Genco, K. T., & Yager, G. G. (1985). Assessing the equivalence of paper-and-pencil vs. computerized tests: Demonstration of a promising methodology. *Computers in Human Behavior, 1,* 265–275.

Woods, K. M., & McNamara, J. R. (1980). Confidentiality: Its effects on interviewee behavior. *Professional Psychology, 11,* 714–720.

Yates, B. T., Wagner, J. L., & Suprenant, L. M. (1997). Recall of health-risky behaviors for the prior 2 or 4 weeks via computerized versus printed questionnaire. *Computers in Human Behavior, 13,* 83–110.

Zerbe, W. J., & Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: A reconception. *Academy of Management Review, 12,* 250–264.

Zickar, M. J., & Robie, C. (1998, April). *Modeling faking good on personality items: An item-level analysis.* Paper presented at the Annual Meeting of the Society of Industrial and Organizational Psychology, Dallas, TX.

Received April 6, 1998
Revision received November 30, 1998
Accepted December 9, 1998 ■