Psychological Methods
2013, Vol. 18, No. 1, 53–70

© 2013 American Psychological Association
1082-989X/13/$12.00 DOI: 10.1037/a0031607

Examination of the Equivalence of Self-Report Survey-Based Paper-and-Pencil and Internet Data Collection Methods

Arne Weigold
Notre Dame College

Ingrid K. Weigold and Elizabeth J. Russell
The University of Akron

Self-report survey-based data collection is increasingly carried out using the Internet, as opposed to the traditional paper-and-pencil method. However, previous research on the equivalence of these methods has yielded inconsistent findings. This may be due to methodological and statistical issues present in much of the literature, such as nonequivalent samples in different conditions due to recruitment, participant self-selection to conditions, and data collection procedures, as well as incomplete or inappropriate statistical procedures for examining equivalence. We conducted 2 studies examining the equivalence of paper-and-pencil and Internet data collection that accounted for these issues. In both studies, we used measures of personality, social desirability, and computer self-efficacy, and, in Study 2, we used personal growth initiative to assess quantitative equivalence (i.e., mean equivalence), qualitative equivalence (i.e., internal consistency and intercorrelations), and auxiliary equivalence (i.e., response rates, missing data, completion time, and comfort completing questionnaires using paper-and-pencil and the Internet). Study 1 investigated the effects of completing surveys via paper-and-pencil or the Internet in both traditional (i.e., lab) and natural (i.e., take-home) settings. Results indicated equivalence across conditions, except for auxiliary equivalence aspects of missing data and completion time. Study 2 examined mailed paper-and-pencil and Internet surveys without contact between experimenter and participants. Results indicated equivalence between conditions, except for auxiliary equivalence aspects of response rate for providing an address and completion time. Overall, the findings show that paper-and-pencil and Internet data collection methods are generally equivalent, particularly for quantitative and qualitative equivalence, with nonequivalence only for some aspects of auxiliary equivalence.

Keywords: equivalence testing, quantitative equivalence, qualitative equivalence, auxiliary equivalence, Internet

As psychological research data collection has moved into the realm of the Internet, data collection has become more convenient in many ways. Compared with traditional data collection methods, such as paper-and-pencil, potential benefits of using the Internet to gather research participants include drawing larger, more diverse samples, accessing difficult-to-reach populations, limiting the costs of survey administration, and eliminating the need for data entry (Buchanan & Smith, 1999; Cantrell & Lupinacci, 2007; Gosling, Vazire, Srivastava, & John, 2004). There is also more flexibility with how survey responses are presented on the Internet (Boyer, Olson, Calantone, & Jackson, 2002). Finally, there are possible benefits for the participants as well, such as comfort, user-friendliness, and convenience (e.g., Naus, Philipp, & Samsi, 2009). Due to these advantages, researchers are increasingly using the Internet as a way to gather participants, particularly for survey-based research (Buchanan, 2007; Gosling et al., 2004). This is hardly surprising, as surveys are one of the most prominent forms of social science research (Berends, 2006) and are easy to adapt for the Internet since they consist of text.

However, there are also potential problems associated with Internet data collection, such as selection bias (Bethlehem, 2010), limiting the sample to those who have Internet access, lack of clarity regarding many aspects of how someone participates (e.g., setting, presence of other people, length of time to complete the study), participant protection, and issues related to not having contact with a researcher (Nosek, Banaji, & Greenwald, 2002). One primary concern is that responses provided via the Internet are not equivalent to those collected using more traditional methods (Buchanan, Ali, et al., 2005). For survey-based designs in particular, this can include both different responses to survey questionnaires (Gosling et al., 2004) and different psychometric properties depending on methodology (Buchanan, 2002).

The popularity of using the Internet to collect data has generated a wealth of research on the comparability of quantitative surveys completed with the standard paper-and-pencil procedure, either in-person or through the mail, or through the Internet. Some researchers have generally concluded that the two methods are equivalent (e.g., Gosling et al., 2004; also see Lewis, Watson, & White, 2009). For example, in one of the seminal articles examining comparability, Buchanan and Smith (1999) developed an

Arne Weigold, Department of Psychology, Notre Dame College; Ingrid K. Weigold and Elizabeth J. Russell, Department of Counseling, The University of Akron.

Correspondence concerning this article should be addressed to Arne Weigold, Department of Psychology, Notre Dame College, 4545 College Road, South Euclid, OH 44121-4293. E-mail: aweigold@ndc.edu

online version of an existing paper-and-pencil self-monitoring questionnaire. They proceeded to collect data via the Internet and compare it to a sample of undergraduate college students who completed the paper-and-pencil version. Based on reliability and confirmatory factor analyses, they concluded that the psychometric properties for the Internet version were at least as good as, if not better than, the paper-and-pencil form. In a second seminal article, Stanton (1998) asked employees to complete either a mailed paper-and-pencil or online packet of questionnaires including demographics, dichotomously scored items related to supervision, and an adapted part of an existing paper-and-pencil organizational justice scale. Mean comparisons suggested no differences in amount of missing data and item variability, and confirmatory factor analysis indicated similar item loadings and covariances.

Although some research has shown the equivalence of paper-and-pencil and Internet methods, a large number of studies have yielded findings that challenge this comparability in a variety of domains, including quantitative item or scale score responses, internal consistency, factor structure, response rates, and/or skipped items (e.g., Barak & Cohen, 2002; Buchanan, Ali, et al., 2005; Carini, Hayek, Kuh, Kennedy, & Ouimet, 2003; Cronk & West, 2002; Davis, 1999; Joinson, 1999; McDonald & Adam, 2003; Naus et al., 2009; Ployhart, Weekley, Holtz, & Kemp, 2003; Sethuraman, Kerin, & Cron, 2005; Shih & Fan, 2008; Whittier, Seeley, & St. Lawrence, 2004). These findings indicate that comparability cannot simply be assumed (e.g., Buchanan, 2003; Buchanan, Ali, et al., 2005; Dolnicar, Laesser, & Matus, 2009). There are two main possible explanations for these findings. First, there may be fundamental differences due to administration via paper-and-pencil or the Internet. For example, questionnaires measuring certain constructs (e.g., sensitive information, computer anxiety, or negative affect) might interact with the method of administration in terms of self-disclosure, computer anxiety, and computer aversion (Schulenberg & Yutrzenka, 1999; see Buchanan, 2003; Wiechmann & Ryan, 2003). Second, there may be methodological or statistical issues present within the studies (Epstein, Klinkenberg, Wiley, & McKinley, 2001; Lewis et al., 2009) that cause the results to inaccurately indicate differences between the two methods of administration. The presence of such issues would make it impossible to determine whether the findings are due to the different methods of administration, methodological or statistical issues, or both.

Taken together, the great variety of findings and underlying methodologies present in the literature raises questions regarding methodological and statistical considerations that limit the usefulness of results from many of the studies currently available on the equivalence of paper-and-pencil and Internet methods (e.g., Epstein et al., 2001; Lewis et al., 2009; see Buchanan, 2003). Therefore, the purpose of the current studies was to examine the equivalence of several forms of paper-and-pencil and Internet survey data collection procedures while avoiding the limitations present in previous research by using an experimental design with randomly assigned participants and control for experimenter effects, as well as suitable statistical analysis methodology (i.e., equivalence testing) of the data. We expected that without the presence of methodological and statistical issues, our results would indicate equivalence.

Methodological Considerations

The main methodological issue associated with research on the comparability of paper-and-pencil and Internet conditions is having nonequivalent samples in different conditions (Epstein et al., 2001; see Buchanan, 2003). This nonequivalence can take several forms, including recruitment procedure differences, participant self-selection to conditions, and inconsistencies in the data collection procedures, such as setting and contact with the researcher. It can affect both studies that are conducted in a laboratory setting and those that include mailed or e-mailed surveys. Although a full review of the literature is beyond the scope of this article, we have attempted to provide several salient examples for each of these considerations.

Recruitment Procedures

Many studies have utilized different recruitment procedures for the paper-and-pencil and Internet conditions, with mixed findings regarding comparability (e.g., Buchanan & Smith, 1999; Cole, Bedeian, & Feild, 2006; Davis, 1999; Hewson & Charlton, 2005; Lewis et al., 2009; McDonald & Adam, 2003; M. A. Smith & Leigh, 1997; Stanton, 1998; Whittier et al., 2004). Whittier et al. (2004) compared paper-and-pencil and Internet responses from self-identified gay and bisexual men on demographics and topics such as substance use and sexual beliefs, attitudes, and behaviors. Paper-and-pencil participants were recruited in-person through gay-oriented businesses, whereas Internet participants were recruited online via gay chat rooms. The authors found differences between conditions on demographics and several topics. Stanton (1998; described above) had organizations distribute the paper-and-pencil packets to their employees but individually contacted the Internet participants at different organizations. Buchanan and Smith's (1999; described above) paper-and-pencil participants came from undergraduate classes at one university, whereas they found their Internet participants via Usenet newsgroups. Similarly, Lewis et al. (2009) conducted an experimental survey study examining drinking and driving attitudes and persuasion in which paper-and-pencil participants were recruited from 1st-year psychology undergraduate students at one university, whereas Internet participants were gathered via the researchers' website (with media advertisement used to increase awareness of the experiment). They found that the two conditions were generally equivalent.

Using different recruitment procedures for the paper-and-pencil and Internet conditions may result in groups that are nonequivalent, creating potential confounds when comparing the two conditions. It should be noted that in some cases, this may be used for an intentional examination of external validity, such as when researchers want to compare a general Internet data collection procedure with the more typical university undergraduate paper-and-pencil recruitment (e.g., Buchanan & Smith, 1999; Lewis et al., 2009). However, the fact remains that it does not allow for a direct comparison of the two conditions. There are some studies that have used within-group designs which, due to the design, have identical recruitment procedures for the conditions examined (e.g., Schulenberg & Yutrzenka, 2001; Vallejo, Mañanes, Comeche, & Díaz, 2008; Whitaker, 2007). However, this does not correct for other methodological and statistical considerations.

Participant Self-Selection to Conditions

Another potential methodological issue related to systematic differences between participants in different conditions occurs when the researcher does not randomly select participants' conditions (e.g., Carini et al., 2003; Dolnicar et al., 2009; B. Smith, Smith, Gray, & Ryan, 2007; Smither, Walker, & Yap, 2004). Carini et al. (2003) used data from the National Survey of Student Engagement. Some of the participating schools required students to complete the survey via the Internet, whereas others gave them a choice between paper-and-pencil and Internet. The authors analyzed Internet participants from both groups, although those in the paper-and-pencil condition had all chosen to complete the study in that format. The authors found significant differences in responses between the conditions, even when controlling for age and gender, such as participants in the Internet condition responding more positively to how much they used computers and related technology. Also, B. Smith et al. (2007) examined results from health-related surveys in participants who chose to complete the study via mailed paper-and-pencil or Internet questionnaires given as part of the Millennium Cohort Questionnaire. The authors found some demographic and health-related differences, such as weight and use of tobacco and alcohol, between the conditions.

Having some or all of the participants self-select their condition can create biases for both paper-and-pencil and Internet conditions (Dolnicar et al., 2009). This is particularly true for surveys examining issues such as computer anxiety; those who choose to participate through the Internet may have less computer anxiety than those who select the paper-and-pencil condition (see Buchanan, 2003). Therefore, unless self-selection bias is a focus of the research (such as Carini et al., 2003, and Dolnicar et al., 2009), participants' self-selection to conditions limits the usability of results.

Data Collection Procedures

A third methodological consideration concerning systematic differences between conditions encompasses differences in extraneous factors associated with data collection procedures, such as setting and physical contact with the researcher (e.g., Buchanan & Smith, 1999; Fouladi, McCarthy, & Moller, 2002; Heath, Lawyer, & Rasmussen, 2007; Hewson & Charlton, 2005; Lewis et al., 2009; Whittier et al., 2004). For example, Hewson and Charlton (2005) compared paper-and-pencil and Internet participants who took a health belief survey. One part of the paper-and-pencil sample completed the survey in class, whereas the other part of the paper-and-pencil sample and the Internet sample completed the survey at their leisure at a place of their choosing. The authors found the factor structure and internal consistencies to be generally equivalent between groups, although mean scores for one of the subscales was significantly lower for the Internet sample.

One particularly problematic procedural difference between conditions is that participants in some conditions have contact with an experimenter whereas those in other conditions do not. Potential effects on research outcomes associated with experimenter presence have long been documented (e.g., Miyazaki & Taylor, 2008; Rosenthal, 1966) and are especially salient for survey questionnaires related to sensitive or socially undesirable information (e.g., Evans, Garcia, Garcia, & Baron, 2003; Richman, Kiesler, Weisband, & Drasgow, 1999). Richman et al. (1999) conducted a meta-analysis comparing interviews, paper-and-pencil assessments, and computerized assessments of social desirability distortion. Regarding paper-and-pencil and computerized surveys, the authors found that the two were generally equivalent. However, when including the presence of an experimenter and the ability to backtrack and change answers, those in the computerized condition provided less socially desirable responses when alone and able to backtrack. The authors concluded that the use of a paper-and-pencil or computer assessment in itself had no significant impact, but that experimenter presence did result in differences between conditions.

Some researchers have directly compared experimenter presence for paper-and-pencil and Internet conditions (e.g., Cronk & West, 2002; Evans et al., 2003; Wood, Nosko, Desmarais, Ross, & Irvine, 2006), with conflicting results. Wood et al. (2006) examined the impact of paper-and-pencil and Internet data collection procedures and experimenter presence on completion of a long survey that included sensitive questionnaires (i.e., containing sexual content). Three conditions were compared: experimenter-present paper-and-pencil, experimenter-present Internet, and experimenter-absent Internet. They found the paper-and-pencil condition contained more missing data than either Internet condition. They did not find an impact of experimenter presence. In contrast, Evans et al. (2003) conducted two studies examining paper-and-pencil and Internet conditions, in addition to experimenter presence and laboratory setting, on European American participants' racial bias distortion. The first study compared three conditions: in-lab, experimenter-present, and paper-and-pencil; in-lab, experimenter-present, and Internet; and out-of-lab, experimenter-absent, and Internet. Participants chose whether they would participate in or out of the lab. The authors found that the first two conditions were significantly more likely to show racial bias distortion compared to the third condition. In the second study, the authors examined whether this effect was due to participants being outside of the lab or away from the experimenter. Three different groups, all Internet-based, were compared: in-lab, experimenter-present; in-lab, experimenter-absent; and out-of-lab, experimenter-absent. Participants were randomly assigned to the conditions. Racial bias distortion was significantly higher for the experimenter-present condition than for the other two. The authors concluded that the presence or absence of the experimenter accounted for the group differences more so than the mode of completion or setting.

Data collection procedural differences between conditions, particularly experimenter presence, can result in confounds. Consequently, the usability of results from studies with these differences will be limited.

Overall, our review of the literature indicates that there are a number of methodological problems, namely differences in recruitment procedures between conditions, participant self-selection to conditions, and data collection procedural differences, which might account for the mixed and often conflicting results regarding comparability of paper-and-pencil and Internet conditions (see Richman et al., 1999). Although some studies have examined the impact of some of these factors, none has addressed all of them in a systematic manner. Thus, research on this topic that takes all these factors into account is needed and is the first major objective of the present research. In addition to these issues, much of the past research has also had issues with statistical analysis methodology, which is the second major objective of the present research.

Statistical Considerations: Equivalence Testing

Types of Equivalence

Researchers have described two main types of equivalence for self-report measures: qualitative and quantitative (Buchanan, 2007; Preckel & Thiemann, 2003; Van de Vijver & Harsveld, 1994). Preckel and Thiemann (2003) defined these terms succinctly: "Whereas qualitative equivalence refers to construct validity, quantitative equivalence mainly addresses the question whether the norm data of one test version can be applied to the other test version" (p. 132). More specifically, qualitative equivalence includes similar internal consistencies, intercorrelations, and/or factor structures (Meyerson & Tryon, 2003), and quantitative equivalence refers to similar mean scores and variances. Researchers have also investigated other aspects of equivalence that go beyond assessing item and test scores, such as response rates and times, missing items, and comfort completing studies using various modalities (e.g., Boyer et al., 2002; Zhang, 2000); however, these have not been systematically defined as corresponding to a particular type of equivalence. Thus, we collectively call these aspects of equivalence auxiliary equivalence.

Many researchers examining paper-and-pencil and Internet conditions do not adequately conduct tests of equivalence but limit the forms of equivalence examined (e.g., Cronk & West, 2002; Davis, 1999; Epstein et al., 2001; Lewis et al., 2009; M. A. Smith & Leigh, 1997; Whittier et al., 2004). Overall, few studies have used all three types of equivalence: quantitative, qualitative, and auxiliary. Even though this might be acceptable for establishing the equivalence of different forms of a specific measure (see Buchanan, 2007; Meyerson & Tryon, 2003), all three types should be used to establish a complete assessment of equivalence.

Analyzing Equivalence

In addition to incompletely examining equivalence, many studies also conduct inappropriate analyses. Whereas the statistics used to analyze qualitative equivalence, such as internal consistency and intercorrelations (Meyerson & Tryon, 2003), are well established and used, analyzing quantitative equivalence and some aspects of auxiliary equivalence is more complicated. A common error is to use null hypothesis significance testing (NHST) to establish equivalence. When applied to group comparisons, NHST refers to the practice of using significance tests to determine whether groups are different. The null hypothesis states that there is no difference between groups (see Nickerson, 2000, for a discussion of various interpretations of the null hypothesis). If the results are statistically significant, that is, the null hypothesis has been rejected, the groups are said to differ. The merits, inappropriate uses, and general controversy surrounding NHST have been thoroughly documented (e.g., Fraley & Marks, 2007; McGrath, 2011; Nickerson, 2000). One of the most frequent ways that NHST is incorrectly used is to conduct equivalence tests (Tryon, 2001; see Rusticus & Lovato, 2011). In this case, two or more groups are compared and, if they are not found to be statistically different, are assumed to be equivalent. Unfortunately, a nonsignificant finding using NHST does not indicate that groups are the same, just that they are not different to a determined degree (Rusticus & Lovato, 2011). Although effect sizes are often added to NHST to provide additional information, they are not sufficient for determining equivalence (see Rogers, Howard, & Vessey, 1993).

Examining the equivalence of group means (mean equivalence testing) takes a different approach to the null and alternate hypotheses. A significant finding indicates the groups are equivalent, whereas a nonsignificant finding indicates differences between groups (Rogers et al., 1993). However, mean equivalence testing is not simply the opposite of NHST. Rather, it examines whether differences between groups, which usually exist since two or more groups are rarely identical, are small enough to be considered inconsequential (Rusticus & Lovato, 2011).

Researchers have suggested that conducting NHST in addition to mean equivalence testing and comparing the results is helpful in providing a more complete picture of equivalence (e.g., Rogers et al., 1993; see Stegner, Bostrom, & Greenfield, 1996). The results of mean equivalence testing and NHST do not always agree (Cribbie, Gruman, & Arpin-Cribbie, 2004). According to Rogers et al. (1993), there are two possible outcomes for each test (rejection or failure to reject the null hypothesis) resulting in four different possible outcomes for the combination of the two tests. If the mean equivalence test is significant and the NHST is not, the groups are considered equivalent. If the mean equivalence test is not significant and the NHST is significant, a difference exists among the compared conditions. If both the mean equivalence test and NHST are significant, there is likely a difference among conditions, although it is small. Finally, if neither the mean equivalence test nor the NHST are significant, the findings are inconclusive, and a determination of equivalence cannot be made. NHST should only be used to provide this type of additional information; it is not suitable for assessing equivalence.

Studies comparing paper-and-pencil and Internet conditions have used mean equivalence testing and NHST to various degrees. A large number of studies relied on NHST without corresponding mean equivalence tests or qualitative equivalence to determine comparability among groups (e.g., Carini et al., 2003; Church, 2001; Cronk & West, 2002; Dolnicar et al., 2009; Heath et al., 2007; Huang, 2006; Joinson, 1999; Kiernan, Kiernan, Oyler, & Giles, 2005; McCabe, 2004; M. A. Smith & Leigh, 1997; Smither et al., 2004; Whittier et al., 2004), whereas others included qualitative equivalence but still only used NHST to assess mean equivalence (e.g., Bandilla, Bosnjak, & Altdorfer, 2003; Barak & Cohen, 2002; Buchanan & Smith, 1999; Davis, 1999; Fouladi et al., 2002; Herrero & Meneses, 2006; Hewson & Charlton, 2005; Naus et al., 2009; Pettit, 2002; Schulenberg & Yutrzenka, 2001; Smither et al., 2004; Stanton, 1998; Vallejo et al., 2008). Interestingly, at least one study has used mean equivalence testing without NHST, making it impossible to know whether any differences found are statistically significant (Epstein et al., 2001).

Only a handful of studies have investigated the comparability of paper-and-pencil and Internet conditions using both mean equivalence testing and NHST. Lewis et al. (2009; described earlier) used both Schuirmann's t test of equivalence (Schuirmann, 1987) and equivalence confidence intervals (Westlake, 1976; see Rogers et al., 1993; Seaman & Serlin, 1998; Tryon & Lewis, 2008) to examine mean equivalence; if conducted correctly, the t test and confidence intervals should reach the same conclusions (Schuirmann, 1987; see Rogers et al., 1993). These tests indicated that most, although not all, comparisons were equivalent. Using a 2 × 2 × 2 analysis of variance (ANOVA) to conduct NHST, the authors found no significant interactions or main effects. The overall pattern of results supported mean equivalence, although the findings were inconclusive for several comparisons, so the authors were unable to determine their equivalence. However, the study had some of the methodological issues described earlier, and it is unclear whether and how the different recruitment procedures for paper-and-pencil and Internet conditions contributed to the results.

In general, our review of the literature indicated that the majority of paper-and-pencil and Internet comparisons did not use complete and/or appropriate statistical analyses to determine equivalence. The present research seeks to completely and correctly examine equivalence, as well as provide an example of how to do so for future studies.

The Current Studies

The purpose of the current studies was to examine the equivalence of paper-and-pencil and Internet data collection for self-report survey-based measures while accounting for several methodological and statistical issues. Study 1 investigated the equivalence of completing the measures via paper-and-pencil or the Internet, as well as doing so in a traditional setting (i.e., in the lab under the supervision of an experimenter) or a natural setting (i.e., at the participants' leisure). Study 2 examined paper-and-pencil condition and Internet condition equivalence without contact between the participants and an experimenter (i.e., through mail and e-mail). Both studies employed several popular measures of specifically chosen constructs (i.e., personality, social desirability, computer self-efficacy; positive psychology in Study 2 only) and four popular demographic questions. We addressed the methodological issues present in much of the literature on paper-and-pencil and Internet equivalence through standardized recruitment, randomized condition assignment, and standardized data collection procedures. We assessed several forms of equivalence using appropriate statistical analyses. We examined quantitative equivalence of the measures using mean equivalence testing. We assessed qualitative equivalence of the measures by inspecting internal consistency and intercorrelations among measure subscales. Finally, we investigated aspects of auxiliary equivalence: response rates and missing data using frequency analysis and completion time and comfort completing questionnaires using mean equivalence testing. We expected the groups to be equivalent for all of these in both studies.

Study 1

Study 1 sought to examine the equivalence of paper-and-pencil and Internet survey conditions while accounting for methodological and statistical issues present in the literature (e.g., Epstein et al., 2001; Lewis et al., 2009). Recruitment procedures were identical for the conditions, and participants were randomly assigned to the conditions. We investigated effects resulting from participants completing the questionnaires in the lab or at their leisure, while holding the other data collection procedures constant. We examined the selected measures' quantitative, qualitative, and auxiliary equivalence.

Method

Participants. Participants included undergraduate students from two schools of higher education. One hundred eighty-one students from a large, urban, Midwestern university participated. These included 120 women (66%) and 61 men (34%) whose ages ranged from 18–55 (M = 21.20, SD = 5.36). The majority identified as European American (n = 141, 78%), followed by African American (n = 25, 14%), biracial/multiracial (n = 4, 2%), and Latin American, Asian American, and other (n = 3, 2% each). Year in school included freshmen (n = 76, 42%), sophomores (n = 37, 20%), juniors (n = 36, 20%), seniors (n = 26, 14%), and other (n = 6, 3%).

Seventy-five students from a small, rural, Midwestern college also participated. There were 55 women (73%) and 20 men (27%) ranging in age from 18–22 (M = 18.67, SD = 0.86). Students identified primarily as European American (n = 67, 89%), followed by African American (n = 5, 7%), Latin American (n = 2, 3%), and Native American (n = 1, 1%). The majority were freshmen (n = 56, 75%), followed by sophomores (n = 15, 20%), juniors (n = 3, 4%), and seniors (n = 1, 1%).

Instruments. We chose to limit the length and number of questionnaires in an effort to avoid participant fatigue. First, we chose a measure of the five-factor model of personality (Costa & McCrae, 1992) because of the prominence of personality questionnaire data collection in psychology in general and via the Internet (Buchanan, 2007; see Kramer, Bernstein, & Phares, 2009). Second, we added social desirability due to its proliferation in computer and paper-and-pencil research, as well as the possible differences between conditions associated with such questionnaires (e.g., Joinson, 1999; Richman et al., 1999). Third, we selected computer self-efficacy due to its potential to influence computer anxiety and performance, possibly resulting in differences between conditions (see Compeau & Higgins, 1995; Wiechmann & Ryan, 2003). Finally, we included four commonly asked demographic questions: gender, age, ethnic/racial identity, and, for student samples, year in school (see American Psychological Association, 2010). It is important to note that the measure for each construct was carefully chosen based on its popularity and brevity; we were not seeking to validate a particular measure for Internet use.

Personality. The International Personality Item Pool (2011 [IPIP]) 50-item measure corresponding to Costa and McCrae's (1992) NEO Personality Inventory—Revised (NEO-PI–R) five-factor model was used to measure personality (IPIP, 2011; see Goldberg, 1999). The IPIP is an international online collaboratory that provides free, unrestricted, instantly accessible measures of personality for research (Goldberg et al., 2006). We selected this particular questionnaire since it has been validated for Internet administration (Buchanan, Johnson, & Goldberg, 2005). The measure consists of 10 items for each of the five factors: Extraversion, Neuroticism, Openness to Experience, Agreeableness, and Conscientiousness. Participants respond to items using a 5-point Likert-type scale ranging from 1 (*Very Inaccurate*) to 5 (*Very Accurate*). In the current study, participants circled a number (paper-and-pencil) or checked a radio button corresponding to a number (Internet) to indicate their answer for each item; the latter is consistent with the Internet validation of the questionnaire (Buchanan, Johnson, & Goldberg, 2005). Items are reverse scored as appropriate and then summed to create the five factors, with higher scores indicating higher levels of the particular personality trait. Internal consistencies ranged from .76 (Openness) to .86 (Neuroticism) for the Internet and .77 (Agreeableness) to .86 (Extraversion and Neuroticism) for paper-and-pencil (Buchanan, Johnson, & Goldberg, 2005). Correlations between each factor and its corresponding NEO-PI–R factor ranged from .70 (Agreeableness) to .82 (Neuroticism; IPIP, 2011), and the factors also correlated in expected directions with behaviors thought to be related to particular aspects of personality (Buchanan, Johnson, & Goldberg, 2005).

Social desirability. The 13-item short form of the Marlowe-Crowne Social Desirability Scale (MCSDS-13; Reynolds, 1982) was selected as a measure of social desirability. This particular measure was chosen because it is a popular shortened form of the original scale (see Barger, 2002). Participants respond to each item as either *true* or *false*. In the current study, participants circled "T" or "F" (paper-and-pencil) or selected "T" or "F" from a drop-down menu (Internet) to indicate their answers. Items are reverse scored as appropriate and summed, with higher scores indicating higher levels of social desirability. The single-factor structure was created by selecting the highest loading items of the original scale. The original study (Reynolds, 1982) reported the internal consistency as .76 and showed that the MCSDS-13 correlated significantly with the original MCSDS and the Edwards Social Desirability Scale (r = .93 and .41, respectively). Zook and Sipps (1985) found adequate test–retest reliability over a 6-week period (r = .74), as well as internal consistency for both men (.72) and women (.75). In addition, Fischer and Fick (1993) found support for internal consistency (.87) and confirmed the factor structure; they also reported a high correlation between the MCSDS-13 and the original MCSDS (r = .97).

Computer self-efficacy. The Computer Self-Efficacy Scale (CSE; Murphy, Coover, & Owen, 1989) was selected to measure computer self-efficacy. We chose this particular measure because of its popularity (see Khorrami-Arani, 2001) and because its three factors correlated significantly and in the expected directions with measures of computer anxiety and computer attitudes (Harrison & Rainer, 1992). The CSE is a 32-item measure of a person's belief in his or her ability to perform tasks on a computer and consists of three subscales corresponding to different skill levels: beginner (CSE Beginner; 16 items), advanced/conceptual (CSE Conceptual; 13 items), and advanced/mainframe (CSE Mainframe; 3 items; Murphy et al., 1989). Participants respond to each item using a 5-point Likert-type scale ranging from 1 (*Very little confidence*) to 5 (*Quite a lot of confidence*). For the current study, participants were asked to write (paper-and-pencil) or type (Internet) their answers beside each statement. Items are then summed to create the subscales. Murphy et al. (1989) provided evidence of the three-factor structure of the CSE, with internal consistency ranging from .92 (CSE Mainframe) to .97 (CSE Beginner). The authors did not specify whether the data were collected via the computer or paper-and-pencil. A later study using mailed paper-and-pencil questionnaires replicated the factor structure (Harrison & Rainer, 1992). Finally, a study using an Internet-based data collection procedure replicated the number of factors, but not the loadings, of the original questionnaire (Barbeite & Weiss, 2004).

Demographics. Participants completed a demographic questionnaire assessing gender, age, ethnic/racial identity, and year in school. In addition, two questions measured comfort completing paper-and-pencil and Internet questionnaires. One asked, "How comfortable are you completing questionnaires using paper-and-pencil?" whereas the other one asked, "How comfortable are you completing questionnaires online using computers?" Participants responded to both questions using a Likert-type scale ranging from 1 (*Not at all comfortable*) to 6 (*Very comfortable*). The two items were not summed.

Completion time. At the beginning and end of the survey were questions regarding the current date and time. These were used to calculate total completion time.

Procedure. Participants from both institutions were recruited through the Department of Psychology subject pools according to the standards and procedures required by each school. Recruitment procedures were identical across conditions; participants were only told that they would be filling out several questionnaires as part of a study. After participants had signed up, they met with a researcher at their school and were randomly assigned to one of four groups: lab paper-and-pencil, lab Internet, take-home paper-and-pencil, or take-home Internet. Participants were not informed that there were different conditions. All data collection procedures were held constant between conditions except for setting and method of completion.

Those in the lab paper-and-pencil condition were given an informed consent form, which they signed and returned to the researcher; they were offered a blank copy to keep. They were then given a packet of questionnaires to complete with the experimenter present and available to answer questions. The three main questionnaires were counterbalanced to prevent order effects; the demographics always followed the final questionnaire. Questionnaires were double-sided, and participants were allowed to backtrack until they handed the packet back to the experimenter. After finishing and returning the packet, participants were given a written debriefing and thanked.

Those in the lab Internet condition underwent identical procedures with the exception that they completed the informed consent, questionnaires, and debriefing on a website accessed via a laptop provided by the experimenter. Participants were required to indicate their informed consent before the website allowed them to proceed to the questionnaires. As they finished one screen, they pressed "Next" to continue to the following one. The questionnaires were again counterbalanced. They were also formatted so that information was consistently displayed in the paper-and-pencil and Internet conditions. For example, Items 1–14 of the personality questionnaire appeared on the first page/screen, Items 15–33 appeared on the second page/screen, and Items 34–50 appeared on the third page/screen. Participants were allowed to backtrack as long as they had not turned in the survey (i.e., pressed "Done" or "Exit the survey"), and they were not required to answer any questions with the exception of indicating informed consent.

Participants in the take-home paper-and-pencil condition were given a packet of questionnaires along with a brief instruction sheet by the experimenter. These packets were identical to those in the lab paper-and-pencil condition, except that they included two copies of the consent form and a completed return envelope. The experimenter discussed the instruction sheet with the participants and answered any questions. Participants were informed they would have one week to complete and return the questionnaires via mail.

Finally, those in the take-home Internet condition experienced identical procedures to those in the take-home paper-and-pencil condition with the exception that they were given a slip of paper with a web link and the experimenter's contact information instead of the questionnaire packet. The web links sent participants to identical questionnaires and procedures as those in the lab Internet condition.

Analyses. Analyses were conducted to examine quantitative, qualitative, and auxiliary equivalence. Since mean equivalence testing is relatively uncommon compared to NHST (see Lewis et al., 2009), we have included a discussion of how to calculate mean equivalence.

Analyses overview. First, we examined the quantitative equivalence of the main variables using mean equivalence tests, adding ANOVA-based NHST to examine the pattern of results. Second, we investigated qualitative equivalence. Internal consistency for the main measures was determined using Cronbach's alpha or the Kuder-Richardson 20, as appropriate. Based on the procedure used by Preckel and Thiemann (2003) for assessing qualitative equivalence, we also compared intercorrelations among the main variables' subscales using χ^2 based on Fisher's r-to-z transformation. Since our social desirability measure did not have subscales, we correlated social desirability with each of the other variables' subscales as qualitative equivalence is also appropriately assessed by examining correlations among diverse measures (Meyerson & Tryon, 2003; Preckel & Thiemann, 2003). Finally, we examined aspects of auxiliary equivalence. We investigated two of these aspects, response rates and missing data, using frequency analyses. Since we were interested in examining the pattern of results, we also conducted chi-square analyses. Similar to the role of NHST in mean equivalence testing, the chi-square analyses were used not to determine equivalence but to provide a more complete picture. We examined the other two aspects of auxiliary equivalence, completion time and comfort completing questionnaires, using mean equivalence tests and added ANOVA-based NHST statistics to examine the pattern of results.

Calculating mean equivalence. We assessed mean equivalence for the main variables, completion time, and comfort completing questionnaires using equivalence confidence intervals. There are two aspects to computing mean equivalence: selecting the equivalence interval around the mean of the reference condition and calculating the confidence interval around the difference of the means of the two conditions being compared (see Rogers et al., 1993). The two conditions are considered to be equivalent only if the latter interval is completely enclosed within the former interval; it is not sufficient for the mean difference between the conditions to be within the equivalence interval.

Our first step in analyzing mean equivalence was to select the equivalence interval. This interval represents how far two conditions can be apart and still be considered equivalent (Rogers et al., 1993). It is important to select an interval that is neither too lenient (i.e., consequential differences between conditions are thought to be negligible) nor too stringent (i.e., trivial differences between conditions indicate the conditions are not comparable). The equivalence interval is computed as a percentage of the reference condition mean (e.g., 20%) that is applied around the mean of the reference condition (e.g., +/−20%). In this +/−20% example, if the reference condition's mean score was 5, the equivalence interval would be +/−1 (or 5 * 0.20). It should be noted that the equivalence interval is based on the observed mean score only and not the reference condition's possible measurement range.

The equivalence interval of +/−20% around the mean of the reference condition is the prevalent interval in the social sciences (Rusticus & Lovato, 2011). It should not be confused with the conventionally accepted risk of committing a Type II error in NHST, which is also 20%. This particular interval has been applied in several studies comparing Internet and other conditions (e.g., Epstein et al., 2001; Lewis et al., 2009; Steele, Mummery, & Dwyer, 2009). It was originally drawn from the general international and U.S. standards for drug bioequivalence testing (Stegner et al., 1996; see Patterson & Jones, 2006).

Although there is precedence for using an equivalence interval of +/−20% around the mean of the reference condition, this number is not without controversy. Unlike the numerical conventions associated with NHST, such as an alpha level of $<.05$ (e.g., Cohen, 1994) or a power level of $\geq.80$ (Cohen, 1992), no generally accepted standard for equivalence intervals yet exists in the social sciences literature (Epstein et al., 2001). This is likely due to the relatively low number of studies employing equivalence testing (see Lewis et al., 2009). Some researchers have selected equivalence intervals based on the topic being examined (e.g., Rogers et al., 1993; Rusticus & Lovato, 2011). For instance, in their three examples of how to conduct equivalence testing, Rogers et al. (1993) employed +/−20% for assessing both baseline equivalence and effect size equivalence in a meta-analysis but +/−10% when examining the equivalence of Minnesota Multiphasic Personality Inventory (MMPI) profiles.

Based on our review of the literature, as well as general considerations, we have selected to use the equivalence interval of +/−20% around the mean of the reference condition. Using measures that are equivalent at this interval is similar to using different measures of the same construct; there will be some acceptable variability in scores between the measures, even if both have strong psychometric properties. However, similar to other numerical conventions in psychology, it should be noted that +/−20% may not be appropriate for all types of psychological research. For example, the standard for an acceptable coefficient alpha of $>.70$ should be raised when important decisions need to be made based on the measure in question (see Drummond & Jones, 2010). Similarly, using equivalence intervals of less than +/−20% may be necessary when small differences in results can have practical consequences, such as in research involving suicide assessments, clinical interventions, or employment evaluations. It remains incumbent upon researchers to thoughtfully select whether to use the +/−20% interval or a different one.

The reference condition for the current study was the lab paper-and-pencil condition; the +/−20% equivalence interval was calculated around the mean for this condition. Epstein et al. (2001) and Lewis et al. (2009) selected their paper-and-pencil conditions as the basis for the equivalence interval as "it represents the traditional approach" (Lewis et al., 2009, p. 111); we chose the lab paper-and-pencil condition for the same reason. Due to the current lack of convention regarding equivalence intervals in psychological research, for comparisons that did not reach equivalence at +/−20%, we examined at what interval the conditions would be considered equivalent.

After selecting the equivalence interval, we created the confidence interval (CI) around the difference of the means of the two conditions being compared (Rogers et al., 1993; Rusticus & Lovato, 2011). If the CI is completely enclosed in the equivalence interval, the two conditions are considered equivalent (Rogers et al., 1993); it is not enough for only the mean difference to be within the equivalence interval. The CI used in equivalence testing is computed according to the generally accepted standards for CIs. The only exception to this is that the CI should be calculated at 100(1 − 2α); the reasons for this are discussed in-depth in Rogers et al. (1993). For example, a standard NHST alpha level of $p \leq .05$ results in a 90% CI. We employed an alpha level of $p \leq .017$ for these particular analyses ($p \leq .05$ with a Bonferroni correction; discussed in the quantitative equivalence section), which resulted in a 96.6% CI calculated around the mean difference between conditions and based on the standard error. Since a CI must be completely enclosed within the equivalence interval in order for the two conditions to be considered equivalent, our use of a smaller alpha level and its correspondingly larger CI made our analyses more conservative. Following the procedure of Rusticus and Lovato (2011), we computed the CIs for the multiple-condition comparisons by conducting a series of one-way ANOVAs using the post hoc Games-Howell procedure. Games-Howell was chosen since it can be used both with unequal *n* sizes and regardless of whether the assumption of homogeneity of variance is met or violated (see Rusticus & Lovato, 2011).

As an example of calculating mean equivalence, we briefly describe one of our findings from Study 1 regarding the equivalence of the lab paper-and-pencil and home paper-and-pencil conditions for CSE Mainframe (discussed in the quantitative equivalence section). First, we calculated the equivalence interval. The mean of the reference condition (i.e., the lab paper-and-pencil condition) was 11.50. Consequently, the +/−20% equivalence interval was +/−2.30 (or 11.50 * 0.20). Second, we calculated the CI. The mean for the home paper-and-pencil condition was 12.30, which was 0.80 above that for the reference condition. The 96.6% CI around the difference between the two means had bounds of −2.34 and 0.74. Since −2.34 was outside of the +/−2.30 range of the equivalence interval, the two conditions were not considered equivalent. It should be noted that if we had inaccurately compared these two conditions by only looking at the mean difference, they would have erroneously been considered equivalent, as the 0.80 difference between the means was within the +/−2.30 equivalence interval.

Results and Discussion

Preliminary analyses. We first examined aspects of quantitative, qualitative, and auxiliary equivalence between the two schools to determine whether they could be combined. The schools were equivalent; we therefore combined them for all analyses. Details regarding these equivalence analyses are available upon request from the first author. The 256 participants were split into the following four conditions: 58 (23%) in lab paper-and-pencil, 60 (23%) in take-home paper-and-pencil, 79 (31%) in lab Internet, and 59 (23%) in take-home Internet. Power analyses for the NHSTs using corrected alpha levels, a medium effect size, and a power level of .80 indicated we met or exceeded the number of participants needed per group for all analyses.

Quantitative equivalence. Means and standard deviations for the main variables are found in Table 1. Mean equivalence tests were conducted on completed questionnaires to determine comparability across conditions for the main variables (see Table 2). Due to the combined number of mean equivalence tests' corresponding NHSTs used here and in the analyses of completion time and comfort completing questionnaires (discussed in the auxiliary equivalence section), we used a Bonferroni corrected alpha level of $p \leq .017$ (.05/3). Consequently, the mean equivalence CIs were set at 96.6%, or 100(1 – 2 * .017). For the main variables, the CIs not enclosed in the 20% equivalence interval included some associated with CSE Conceptual (lab paper-and-pencil compared to the other three conditions), CSE Mainframe (the paper-and-pencil conditions compared to each other), and Neuroticism (both paper-and-pencil conditions compared to take-home Internet). These CIs

Table 1
Study 1 Means, Standard Deviations, and Internal Consistency for the Main Variables, Completion Time, and Comfort Completing Questionnaires

	Lab PP			Home PP			Lab Internet			Home Internet		
Variable	*n*	*M* (*SD*)	α	*n*	*M* (*SD*)	α	*n*	*M* (*SD*)	α	*n*	*M* (*SD*)	α
Extraversion	57	34.58 (6.36)	.82	58	36.45 (7.71)	.89	76	35.96 (6.38)	.84	56	35.96 (6.27)	.84
Neuroticism	57	26.12 (6.51)	.81	60	25.77 (7.91)	.86	75	25.25 (6.67)	.82	59	23.81 (6.62)	.85
Openness	58	35.50 (5.92)	.74	59	37.32 (6.88)	.81	75	34.56 (6.92)	.82	58	36.19 (5.96)	.76
Agreeableness[a]	57	36.11 (5.73)	.80	60	36.33 (5.36)	.76	73	38.22 (4.10)	.63	57	38.60 (5.08)	.76
Conscientiousness[a]	57	36.88 (4.09)	.60	59	34.61 (6.17)	.80	74	35.91 (4.91)	.68	57	36.40 (5.57)	.81
Social Desirability	58	19.41 (2.66)	.62	60	18.18 (2.87)	.69	76	19.07 (2.84)	.67	59	19.46 (2.59)	.63
CSE Beginner	58	64.45 (14.38)	.96	60	69.62 (9.19)	.92	78	67.71 (13.01)	.95	58	67.81 (11.45)	.94
CSE Conceptual	58	42.59 (11.27)	.92	60	46.43 (10.37)	.91	79	46.73 (11.64)	.94	58	46.12 (10.76)	.93
CSE Mainframe	58	11.50 (3.22)	.84	60	12.30 (2.83)	.92	78	12.06 (3.36)	.90	58	11.98 (2.91)	.87
Completion time	58	10.29 (3.99)	—	59	13.76 (7.15)	—	79	10.78 (2.63)	—	57	9.98 (3.72)	—
Comfort PP	58	5.67 (0.54)	—	60	5.82 (0.43)	—	79	5.49 (0.78)	—	58	5.54 (0.86)	—
Comfort Internet	58	5.24 (1.03)	—	60	5.52 (0.97)	—	79	5.47 (0.88)	—	58	5.63 (0.64)	—

Note. Completion time does not include two outliers (one take-home paper-and-pencil, one take-home Internet). Nonapplicable alpha internal consistency estimates are marked with a dash (—). Lab PP = lab paper-and-pencil; Home PP = take-home paper-and-pencil; Home Internet = take-home Internet; CSE = Computer Self-Efficacy Scale; Comfort PP = comfort completing questionnaires using paper-and-pencil; Comfort Internet = comfort completing questionnaires using the Internet.
[a] Alpha levels differ by at least +/−.10.

Table 2
Study 1 Mean Equivalence Tests for the Main Variables, Completion Time, and Comfort Completing Questionnaires

		Lab PP–Home PP 96.6% CI		Lab PP–Lab Internet 96.6% CI		Lab PP–Home Internet 96.6% CI		Home PP–Lab Internet 96.6% CI		Home PP–Home Internet 96.6% CI		Lab Internet–Home Internet 96.6% CI	
Variable	20% interval	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Extraversion	+/−6.92	−5.50	1.76[a]	−4.45	1.69[a]	−4.66	1.89[a]	−2.96	3.93[a]	−3.14	4.11[a]	−3.07	3.06[a]
Neuroticism	+/−5.22	−3.33	4.04[a]	−2.31	4.05[a]	−1.05	5.67	−3.01	4.04[a]	−1.73	5.64	−1.74	4.62[a]
Openness	+/−7.10	−5.09	1.45[a]	−2.13	4.01[a]	−3.73	2.35[a]	−0.54	6.06[a]	−2.15	4.41[a]	−4.71	1.45[a]
Agreeableness	+/−7.22	−3.06	2.60[a]	−4.59	0.37[a]	−5.29	0.30[a]	−4.21	0.44[a]	−4.92	0.40[a]	−2.66	1.90[a]
Conscientiousness	+/−7.38	−0.41	4.94[a]	−1.19	3.14[a]	−2.05	3.00[a]	−4.10	1.42[a]	−4.80	1.21[a]	−3.07	2.07[a]
Social Desirability	+/−3.88	−0.17	2.63[a]	−0.96	1.66[a]	−1.38	1.29[a]	−2.24	0.47[a]	−2.65	0.10[a]	−1.68	0.90[a]
CSE Beginner	+/−12.89	−11.33	0.99[a]	−9.85	3.33[a]	−10.02	3.29[a]	−3.29	7.11[a]	−3.47	7.09[a]	−5.89	5.68[a]
CSE Conceptual	+/−8.52	−9.34	1.65	−9.58	1.29	−9.17	2.10	−5.44	4.85[a]	−5.05	5.67[a]	−4.69	5.90[a]
CSE Mainframe	+/−2.30	−2.34	0.74	−2.13	1.00[a]	−2.05	1.09[a]	−1.21	1.69[a]	−1.14	1.77[a]	−1.40	1.56[a]
Completion time[b]	+/−2.06	−6.42	0.52	−2.16	1.17	−1.66	2.28	0.26	5.07	0.87	6.69	−0.78	2.38
Comfort PP[b]	+/−1.13	−0.39	0.11[a]	−0.13	0.49[a]	−0.24	0.50[a]	0.04	0.61[a]	−0.07	0.62[a]	−0.44	0.34[a]
Comfort Internet	+/−1.05	−0.78	0.23[a]	−0.69	0.23[a]	−0.83	0.05[a]	−0.39	0.49[a]	−0.52	0.30[a]	−0.51	0.20[a]

Note. The equivalence interval is based on the lab paper-and-pencil condition mean. Completion time does not include two outliers (one take-home paper-and-pencil, one take-home Internet). Lab PP = lab paper-and-pencil; Home PP = take-home paper-and-pencil; Home Internet = take-home Internet; CSE = Computer Self-Efficacy Scale; Comfort PP = comfort completing paper-and-pencil questionnaires; Comfort Internet = comfort completing questionnaires using the Internet; CI = confidence interval.
[a] Equivalent at 20%. The upper CI is marked for each significant pair. [b] Variances not equal across conditions.

were equivalent at intervals between 22%–25%. A 2 (mode: paper-and-pencil vs. Internet) × 2 (setting: lab vs. take-home) multivariate analysis of variance (MANOVA) was conducted as an examination of NHST results. Levene's test was not significant at $p \leq .017$ for any of the variables. The interaction was not significant, $F(9, 210) = 1.54$, *ns*, $\eta_p^2 = .06$, nor were the main effects of mode, $F(9, 210) = 1.80$, *ns*, $\eta_p^2 = .07$, or setting, $F(9, 210) = 1.42$, *ns*, $\eta_p^2 = .06$. Therefore, NHST indicated no differences among the groups. Our findings indicated that groups were generally equivalent regarding mean differences, although a few of the comparisons were inconclusive at the 20% interval.

Qualitative equivalence. Qualitative equivalence was assessed using internal consistency and intercorrelations comparisons. The analyses were conducted with the main variables.

Internal consistency. Internal consistency estimates for the nine main variables are shown in Table 1. The majority of alpha levels were ≥.70, indicating good internal consistency, although several were in the marginal range between .60 and .69: Agreeableness lab Internet, Conscientiousness lab paper-and-pencil and lab Internet, and all social desirability conditions. Following the method of Meyerson and Tryon (2003), we defined internal consistency equivalence as alpha levels that differ by <+/−.10. Most of the variables showed equivalent alpha levels, with the exception of Agreeableness and Conscientiousness. For Agreeableness, the lab Internet condition (.63) was significantly lower than the other three conditions (range = .76–.80). Regarding Conscientiousness, the lab paper-and-pencil and lab Internet conditions (.60 and .68, respectively) were both significantly lower than the take-home paper-and-pencil and take-home Internet conditions (.80 and .81, respectively). These findings indicated equivalence for most of the variables, with the exceptions of Agreeableness and Conscientiousness.

Intercorrelations. Intercorrelations were run for the five personality factors and three CSE subscales, as well as social desirability with the personality factors and CSE subscales. We followed the method of Preckel and Thiemann (2003), who used Fisher's *r*-to-*z* transformation to compare the correlations. Since the largest number of comparisons in which any one scale was used was eight, we used a Bonferroni corrected alpha level of $p \leq .006$ (.05/8). None of the comparisons was significantly different across groups, supporting comparability (see Table 3). The magnitudes and directions of the correlations were generally consistent with previous correlation estimates available in the literature (e.g., Bäckström, Björklund, & Larsson, 2009; Harrison & Rainer, 1992; van der Linden, te Nijenhuis, & Bakker, 2010).

Auxiliary equivalence. Auxiliary equivalence included response rates, missing data, completion time, and comfort completing questionnaires. The first two were examined using frequency analyses, and the remaining two were assessed using mean equivalence testing.

Response rates. One person (2%) in the take-home paper-and-pencil condition did not return the survey. In addition, one person (1%) in the lab Internet and two people (3%) in the take-home Internet condition started but did not complete the study. Thus, response rates appeared to be equivalent. The chi-square analysis did not indicate a statistically significant difference among the groups, $\chi^2(3) = 2.17$, *ns*.

Missing data. We examined missing data for the demographic and main variables (there was no missing data for completion time or the two comfort questions). The examination of NHST included two chi-square analyses; we used an alpha correction of $p \leq .025$ (.05/2). Of the four demographic questions (gender, age, ethnic/racial identity, and year in school), only two participants (3%) in the lab paper-and-pencil condition were missing data on ethnic/racial identity, which suggested equivalence across conditions. A chi-square analysis indicated that the difference was not statistically significant, $\chi^2(3) = 6.88$, *ns*.

An examination of the main variables' data points indicated missing data in all conditions. However, the number of participants missing data in each condition varied: four (7%) for lab paper-

Table 3
Study 1 Comparison of Correlations Across Conditions for the Personality, Social Desirability, and Computer Self-Efficacy Factors

Variable 1	Variable 2	Lab PP r (n)	Home PP r (n)	Lab Internet r (n)	Home Internet r (n)	r-to-z χ^2
Extraversion	Neuroticism	−.35 (56)	−.41 (58)	−.42 (72)	−.55 (56)	1.84
Extraversion	Openness	.05 (57)	−.08 (57)	.21 (74)	.05 (55)	2.72
Extraversion	Agreeableness	.28 (56)	.30 (58)	.19 (72)	.23 (55)	0.51
Extraversion	Conscientiousness	−.12 (56)	.29 (58)	.19 (71)	.13 (54)	5.15
Neuroticism	Openness	.02 (57)	.20 (59)	.08 (71)	.01 (58)	1.31
Neuroticism	Agreeableness	−.44 (56)	−.32 (60)	−.25 (70)	−.41 (57)	1.75
Neuroticism	Conscientiousness	−.05 (56)	−.43 (59)	−.34 (70)	−.14 (57)	5.93
Openness	Agreeableness	−.03 (57)	.10 (59)	.15 (70)	.10 (56)	1.03
Openness	Conscientiousness	.12 (57)	−.25 (58)	.07 (70)	.10 (56)	5.21
Agreeableness	Conscientiousness	.04 (56)	.18 (59)	.46 (69)	.26 (55)	6.62
Social Desirability	Extraversion	−.18 (57)	.36 (58)	.19 (73)	.29 (56)	9.95
Social Desirability	Neuroticism	−.02 (57)	−.38 (60)	−.37 (72)	−.28 (59)	5.28
Social Desirability	Openness	−.00 (58)	.07 (59)	.16 (72)	−.03 (58)	1.36
Social Desirability	Agreeableness	.11 (57)	.46 (60)	.41 (71)	.46 (57)	5.64
Social Desirability	Conscientiousness	.03 (57)	.35 (59)	.37 (71)	.22 (57)	4.67
Social Desirability	CSE Beginner	−.08 (58)	.20 (60)	.15 (75)	−.13 (58)	4.79
Social Desirability	CSE Conceptual	.14 (58)	.33 (60)	.13 (76)	−.10 (58)	5.51
Social Desirability	CSE Mainframe	.07 (58)	.21 (60)	.10 (75)	−.17 (58)	4.41
CSE Beginner	CSE Conceptual	.81 (58)	.81 (60)	.81 (78)	.84 (57)	0.37
CSE Beginner	CSE Mainframe	.76 (58)	.63 (60)	.69 (77)	.59 (57)	3.21
CSE Conceptual	CSE Mainframe	.58 (58)	.54 (60)	.64 (78)	.59 (57)	0.80

Note. Comparisons were made using χ^2 analyses based on Fisher's r-to-z transformation; none were significant at the corrected alpha level of $p \leq .006$. Lab PP = lab paper-and-pencil; Home PP = take-home paper-and-pencil; Home Internet = take-home Internet; CSE = Computer Self-Efficacy Scale.

and-pencil, three (5%) for take-home paper-and-pencil, 18 (23%) for lab Internet, and nine (15%) for take-home Internet. These suggested that more participants in the Internet conditions were missing data than in the paper-and-pencil conditions. The overall chi-square analysis was statistically significant, $\chi^2(3) = 12.02$, $p < .01$. We further investigated the differences among conditions by examining the amount of missing data. Across conditions, the number of missing data points was low. Out of the 34 participants with missing data, the majority ($n = 24$, 68%) were only missing one item, with some missing two ($n = 8$, 24%) or three ($n = 2$, 6%) items, out of the potential 95 data points per person. Taken together, these findings indicated that although there were differences in the presence of missing data across conditions, the actual amount of missing data was trivial.

Completion time. Means and standard deviations for completion time are found in Table 1. We used mean equivalence testing to compare groups. Examination of the frequencies of completion time yielded two outliers, one in the take-home paper-and-pencil condition (182 min, or approximately 3 hr) and one in the take-home Internet condition (596 min, or approximately 10 hr). These outliers were removed for the mean equivalence test. Equivalence 96.6% CIs indicated that none of the groups was equivalent (see Table 2). Comparisons involving the take-home paper-and-pencil condition were equivalent at intervals between 50%–66%, and comparisons not involving this condition were equivalent between 22%–24%. A 2 (mode) × 2 (setting) ANOVA was conducted for NHST. Levene's test was significant, $F(3, 250) = 11.46$, $p < .001$. The interaction was significant at $p \leq .017$, $F(1, 250) = 13.69$, $p < .001$, $\eta_p^2 = .05$, as was the main effect of mode, $F(1, 250) = 8.11$, $p < .01$, $\eta_p^2 = .03$. The main effect of setting was not significant, $F(1, 250) = 5.34$, *ns*, $\eta_p^2 = .02$. Post hoc analyses indicated the take-home paper-and-pencil condition took significantly more time than the other three conditions. Overall, results indicated that average completion time was not equivalent across conditions.

Comfort completing questionnaires. Means and standard deviations for the two comfort questions—comfort completing questionnaires using paper-and-pencil and the Internet—are shown in Table 1. Equivalence 96.6% CIs suggested comparability for all comparisons (see Table 2). We next conducted a 2 (mode) × 2 (setting) MANOVA on the two questions. Levene's test was significant for paper-and-pencil comfort, $F(3, 252) = 10.91$, $p < .001$. There was a significant main effect of mode, $F(2, 251) = 7.20$, $p = .001$, $\eta_p^2 = .05$, with those in the paper-and-pencil conditions indicating higher levels of comfort completing paper-and-pencil questionnaires than those in the Internet conditions. The interaction was not significant, $F(2, 251) = 0.81$, *ns*, $\eta_p^2 < .01$, nor was the main effect of setting, $F(2, 251) = 1.95$, *ns*, $\eta_p^2 = .02$, at $p \leq .017$. This pattern of results indicated that a difference likely exists between the paper-and-pencil and Internet conditions, although it is not large enough for the groups to not be equivalent.

Overall, our results indicated quantitative and qualitative equivalence for the main variables, as well as comparability for two of the four aspects of auxiliary equivalence (i.e., response rates and comfort completing questionnaires). The conditions differed in regard to the presence of missing data, with more participants in the Internet conditions having missing data points than those in the paper-and-pencil conditions; however, the actual amount of missing data was trivial. There were also differences in total completion time for all condition comparisons, with those in the take-home paper-and-pencil condition taking longer than those in the other conditions.

In Study 1, we held contact with the researcher constant for all conditions in order to demonstrate the equivalence of survey administration for different modes (Internet and paper-and-pencil) and settings (lab and take-home). However, many data collections take place with mailed surveys or e-mailed web links, in which participants do not meet the researcher. Therefore, we conducted Study 2 to examine equivalence of Internet and paper-and-pencil survey administration for when this is the case.

Study 2

Study 2 was similar to Study 1 in general purpose, method, and analysis. However, we only examined two conditions in this study: the mailed paper-and-pencil and the e-mailed Internet conditions, which are two common forms of survey data collection in which the participants do not come to a lab or have any face-to-face contact with the researchers. We also included one additional measure that is often administered without direct contact between the participants and the experimenter.

Method

Participants. Participants were students in undergraduate courses from the same two schools of higher education as in Study 1. One hundred thirty students from the university provided their contact information and returned the survey. There were 102 women (79%) and 28 men (22%), and ages ranged from 18–59 (M = 22.41, SD = 5.65). The majority identified as European American (n = 104, 81%), followed by African American (n = 14, 11%), biracial/multiracial (n = 6, 5%), Asian American and Latin American (n = 2, 2% each), and other (n = 1, 1%). Students identified as freshmen (n = 29, 23%), sophomores (n = 22, 17%), juniors (n = 34, 27%), seniors (n = 39, 31%), and other (n = 4, 3%).

Seventy-three students from the college also provided their contact information and returned the survey. These included 54 women (74%) and 19 men (26%) who ranged in age from 18–22 (M = 19.07, SD = 1.16). The majority identified as European American (n = 60, 82%), followed by Asian American and other (n = 4, 6% each), African American and biracial/multiracial (n = 2, 3% each), and Native American (n = 1, 1%). Most identified as freshmen (n = 46, 63%), followed by sophomores (n = 15, 21%), juniors (n = 6, 8%), seniors (n = 5, 7%), and other (n = 1, 1%).

Instruments. Instruments were mostly identical to those used in Study 1 (i.e., the measures of personality, social desirability, computer self-efficacy, and demographics). However, we added a positive psychology measure: personal growth initiative. We included a positive psychology construct due to positive psychology's increasing popularity in psychology (Diener, 2009) and delivery over the Internet (see Seligman, Steen, Park, & Peterson, 2005; Voiskounsky, 2008).

Personal growth initiative. The Personal Growth Initiative Scale–II (PGIS-II; Robitschek et al., 2012) was used to assess personal growth initiative. Although the measure is new to the literature, and the majority of the psychometric properties were examined in college students using lab paper-and-pencil data collection procedures, it was chosen because confirmatory factor analyses were also conducted on college student and community samples who completed the questionnaire on the Internet. The PGIS-II is a 16-item scale assessing the active and intentional desire to grow as a person. It consists of four factors: Readiness for Change (four items measuring a person's ability to recognize when he or she is ready to start the change process), Planfulness (five items assessing one's ability to make plans regarding the change process), Intentional Behavior (six items assessing one's ability to engage in actual change behaviors), and Using Resources (three items measuring a person's use of external resources during the change process). Participants respond to items using a 6-point Likert-type scale, with responses ranging from 0 (*Disagree Strongly*) to 5 (*Agree Strongly*). In the current study, participants circled a number (paper-and-pencil) or checked a radio button (Internet) to indicate their response. Scores are summed and then averaged, with higher scores indicating higher levels of each factor. Robitschek et al. (2012) has shown evidence for internal consistency, with estimates ranging from .73 (Using Resources) to .91 (Planfulness and Intentional Behavior), as well as adequate test–retest reliability for up to a 6-week period. The factors generally correlated significantly and in the expected directions with the original Personal Growth Initiative Scale and measures of assertiveness and instrumentality; the factors did not correlate significantly with social desirability (Robitschek et al., 2012).

Procedure. As in Study 1, we recruited participants from both institutions through their Department of Psychology subject pools. Recruitment was identical between conditions; participants were not made aware that there were different conditions. After participants signed up for the study, they were provided with a web link randomly sending them to the paper-and-pencil or Internet condition. We chose to have all participants go to an Internet sign-up site initially in order to keep recruitment procedures constant (i.e., everyone had to provide their addresses the same way). All data collection procedures were the same between conditions except for the type of address participants were asked to provide and the method of completing the questionnaires. All participants read and agreed to the informed consent, which described the type of data collection in which they would participate. After agreeing to participate, those in the paper-and-pencil condition were asked to provide a mailing address to which the questionnaires could be mailed, whereas those in the Internet condition were requested to provide an e-mail address to which the survey web link could be e-mailed. All participants received an automatic notice that their information had been received, told they would be sent the survey within several days, and provided with researcher contact information.

Participants in both conditions were sent the survey to complete. The main questionnaires were counterbalanced, with the demographics always following the final questionnaire. The survey began and ended with questions regarding the current date and time. The questionnaires were formatted so that the information was presented consistently in both conditions. Those in the paper-and-pencil condition were mailed an instruction sheet, the questionnaires, a debriefing, and a completed return envelope. The questionnaires were double-sided, and participants were allowed to backtrack. Participants were given several days from when they received the questionnaires to complete and return them. Those in the Internet condition were e-mailed a web link to the questionnaires and debriefing. Participants were allowed to backtrack as long as they had not turned in the survey (i.e., pressed "Done" or "Exit the survey"), and they were not required to answer any questions to move on. They were also given several days from when they received the link to complete the questionnaires.

Analyses. We conducted the same analyses as in Study 1 to examine quantitative, qualitative, and auxiliary equivalence. Since we were comparing two, rather than multiple, groups, we used a series of *t* tests rather than a series of ANOVA, to compute the mean equivalence test CIs and followed the general procedure for calculating two-group comparisons described in Rogers et al. (1993). We selected the specific CIs based on whether or not equal variances were assumed.

Results

Preliminary analyses. As in Study 1, we compared the two schools on aspects of quantitative, qualitative, and auxiliary equivalence, found them to be equivalent, and combined the two samples; details of these analyses are available upon request from the first author. The 203 participants were split into two conditions: 86 (42%) in the paper-and-pencil condition and 117 (58%) in the Internet conditions. Power analyses for the NHSTs using corrected alpha levels, a medium effect size, and a power level of .80 indicated we met or exceeded the number of participants needed per group for all analyses except for one group of the mean equivalence NHST ANOVA for completion time, which reached significance.

Quantitative equivalence. Means and standard deviations for the main variables are shown in Table 4. Mean equivalence tests were conducted to determine the comparability between conditions for the main variables (see Table 5). Due to the combined number of mean equivalence tests' corresponding NHSTs used here and in the analyses of completion time and comfort completing questionnaires (discussed in the auxiliary equivalence section), we used a Bonferroni corrected alpha level of $p \leq .017$ (.05/3); mean equivalence confidence intervals were consequently set at 96.6%, or 100(1 − 2 * .017). Mean equivalence tests indicated that all variables were equivalent between conditions. We then conducted one MANOVA as it is the most appropriate statistic for NHST. Levene's test was not significant for any of the variables. The MANOVA was also not significant, $F(13, 157) = 1.21$, *ns*, $\eta_p^2 = .09$. These findings indicated all variables were equivalent between conditions.

Table 4
Study 2 Means, Standard Deviations, and Internal Consistency for the Main Variables, Completion Time, and Comfort Completing Questionnaires

	Paper-and-pencil			Internet		
Variable	*n*	*M* (*SD*)	α	*n*	*M* (*SD*)	α
Extraversion	83	35.30 (7.98)	.90	112	34.35 (7.19)	.88
Neuroticism	84	26.68 (7.92)	.86	115	25.35 (7.13)	.86
Openness	85	36.86 (5.79)	.72	115	35.88 (5.69)	.73
Agreeableness	86	37.90 (4.73)	.71	112	36.58 (5.12)	.76
Conscientiousness	86	36.41 (6.28)	.84	114	36.92 (5.75)	.83
Social Desirability	86	18.51 (2.80)	.68	114	18.89 (2.88)	.69
CSE Beginner	85	70.21 (9.44)	.92	112	68.41 (10.79)	.93
CSE Conceptual	86	45.19 (10.27)	.91	111	45.73 (11.75)	.93
CSE Mainframe	86	11.72 (3.54)	.90	116	11.54 (3.23)	.90
PGIS RC	84	3.69 (0.84)	.86	117	3.60 (0.85)	.85
PGIS PL	86	3.60 (0.90)	.90	117	3.68 (0.83)	.89
PGIS IB	85	3.76 (0.82)	.87	116	3.75 (0.84)	.85
PGIS UR[a]	85	2.90 (1.23)	.88	116	2.96 (1.06)	.78
Completion time	79	19.11 (10.02)	—	115	12.29 (5.34)	—
Comfort PP	86	5.65 (0.85)	—	116	5.23 (1.18)	—
Comfort Internet	86	5.51 (0.79)	—	117	5.38 (1.06)	—

Note. Completion time does not include four outliers (three paper-and-pencil, one Internet). Nonapplicable internal consistency estimates are marked with a dash (—). CSE = Computer Self-Efficacy Scale; PGIS = Personal Growth Initiative Scale; RC = Readiness for Change; PL = Planfulness; IB = Intentional Behavior; UR = Using Resources; Comfort PP = comfort completing questionnaires using paper-and-pencil; Comfort Internet = comfort completing questionnaires using the Internet.
[a] Alpha levels differ by at least +/−.10.

Table 5
Study 2 Mean Equivalence Tests for the Main Variables, Completion Time, and Comfort Completing Questionnaires

Variable	20% interval	Lower 96.6% CI	Upper 96.6% CI
Extraversion	+/−7.06	−1.38	3.28[a]
Neuroticism	+/−5.34	−0.86	3.72[a]
Openness	+/−7.58	−0.77	2.73[a]
Agreeableness	+/−7.60	−0.20	2.83[a]
Conscientiousness	+/−7.28	−2.34	1.31[a]
Social Desirability	+/−3.70	−1.24	0.49[a]
CSE Beginner	+/−14.04	−1.34	4.94[a]
CSE Conceptual	+/−9.04	−3.96	2.87[a]
CSE Mainframe	+/−2.34	−0.85	1.20[a]
PGIS Readiness for Change	+/−0.74	−0.17	0.34[a]
PGIS Planfulness	+/−0.72	−0.35	0.18[a]
PGIS Intentional Behavior	+/−0.75	−0.24	0.27[a]
PGIS Using Resources	+/−0.58	−0.41	0.28[a]
Completion time[b]	+/−3.82	4.18	9.47
Comfort PP[b]	+/−1.13	−0.11	0.72[a]
Comfort Internet	+/−1.10	−0.15	0.42[a]

Note. The equivalence interval is based on the paper-and-pencil condition mean. Completion time does not include four outliers (three paper-and-pencil, one Internet). CSE = Computer Self-Efficacy Scale; PGIS = Personal Growth Initiative Scale; Comfort PP = comfort completing paper-and-pencil questionnaires; Comfort Internet = comfort completing questionnaires using the Internet; CI = confidence interval.
[a] Equivalent at 20%. The upper CI is marked for each significant pair. [b] Variances not equal between conditions.

Qualitative equivalence. The qualitative equivalence analyses were identical to those used in Study 1. They consisted of examining internal consistencies for and intercorrelations among the main variables.

Internal consistency. Internal consistencies for the 13 main variables are shown in Table 4. Most of the variables had alpha levels at ≥.70, with the exception of social desirability for both the paper-and-pencil and Internet conditions (.68 and .69, respectively). Using Resources differed at .10 (Meyerson & Tryon, 2003) between conditions (paper-and-pencil α = .88 and Internet α = .78). All of the other variables were equivalent, with alpha levels that differed by < +/−.10. These findings indicated that internal consistency was equivalent between conditions for all but one variable.

Intercorrelations. We examined intercorrelations among the five personality factors, the three CSE subscales, and the four PGIS-II subscales. We also correlated social desirability with the other 12 variables (see Table 6). As in Study 1, we followed the method of Preckel and Thiemann (2003), who used Fisher's *r*-to-*z* transformation to compare the correlations. The largest number of comparisons in which any one scale was used was 12, so we used a Bonferroni corrected alpha level of $p \leq .004$ (.05/12). Only the correlations between Neuroticism and Openness (paper-and-pencil $r = -.30$ and Internet $r = .12$) were significantly different from

Table 6
Study 2 Comparisons of Correlations Among Personality, Social Desirability, Computer Self-Efficacy, and Personal Growth Initiative Factors

Variable 1	Variable 2	Paper-and-pencil *r* (*n*)	Internet *r* (*n*)	*r*-to-*z* χ^2	Variable 1	Variable 2	Paper-and-pencil *r* (*n*)	Internet *r* (*n*)	*r*-to-*z* χ^2
Extraversion	Neuroticism	−.36 (81)	−.41 (112)	0.16	Social Desirability	CSE Conceptual	.10 (86)	.29 (108)	1.82
Extraversion	Openness	.22 (82)	.12 (111)	0.48	Social Desirability	CSE Mainframe	.03 (86)	.10 (113)	0.23
Extraversion	Agreeableness	.24 (83)	.14 (108)	0.49	Social Desirability	PGIS RC	.20 (84)	.23 (114)	0.02
Extraversion	Conscientiousness	.24 (83)	.29 (110)	0.13	Social Desirability	PGIS PL	.22 (86)	.33 (114)	0.67
Neuroticism	Openness	−.30 (83)	.12 (114)	8.60[a]	Social Desirability	PGIS IB	.32 (85)	.20 (113)	0.67
Neuroticism	Agreeableness	−.40 (84)	−.27 (111)	0.99	Social Desirability	PGIS UR	.11 (85)	.13 (113)	0.04
Neuroticism	Conscientiousness	−.35 (84)	−.31 (113)	0.09	CSE Beginner	CSE Conceptual	.75 (85)	.77 (108)	0.10
Openness	Agreeableness	−.05 (85)	.10 (111)	1.05	CSE Beginner	CSE Mainframe	.60 (85)	.54 (112)	0.37
Openness	Conscientiousness	.04 (85)	.12 (113)	0.30	CSE Conceptual	CSE Mainframe	.43 (86)	.52 (110)	0.63
Agreeableness	Conscientiousness	.15 (86)	.32 (111)	1.53	PGIS RC	PGIS PL	.71 (84)	.76 (117)	0.56
Social Desirability	Extraversion	.18 (83)	.14 (109)	0.08	PGIS RC	PGIS IB	.60 (84)	.72 (116)	2.17
Social Desirability	Neuroticism	−.29 (84)	−.39 (112)	0.60	PGIS RC	PGIS UR	.50 (83)	.59 (116)	0.77
Social Desirability	Openness	−.03 (85)	.00 (112)	0.04	PGIS PL	PGIS IB	.58 (85)	.79 (116)	7.95
Social Desirability	Agreeableness	.45 (86)	.44 (109)	0.01	PGIS PL	PGIS UR	.34 (85)	.53 (116)	2.65
Social Desirability	Conscientiousness	.26 (86)	.29 (111)	0.05	PGIS IB	PGIS UR	.47 (84)	.49 (115)	0.03
Social Desirability	CSE Beginner	.05 (85)	.10 (109)	0.12					

Note. Comparisons were made using Fisher's *r*-to-*z* transformation. CSE = Computer Self-Efficacy Scale; PGIS = Personal Growth Initiative Scale; RC = Readiness for Change; PL = Planfulness; IB = Intentional Behavior; UR = Using Resources.

[a] Significant at the corrected alpha level of $p \leq .004$.

each other. As in Study 1, the correlation magnitudes and directions were mostly consistent with those found in previous research (e.g., Bäckström et al., 2009; Harrison & Rainer, 1992; van der Linden et al., 2010), with the exception of the personal growth initiative factors and social desirability, which were primarily in the opposite direction of, and somewhat larger than, those reported in the literature (Robitschek et al., 2012).

Auxiliary equivalence. The examination of auxiliary equivalence was identical to that in Study 1. It consisted of assessing response rates and missing data using frequency analyses and assessing completion time and comfort completing questionnaires using mean equivalence testing.

Response rates. Examination of response rates included two chi-square analyses; we used an alpha correction of $p \leq .025$ (.05/2). We investigated response rates on two levels. First, we examined how many participants who agreed to participate on the consent form subsequently provided a valid address. For the paper-and-pencil condition, only 123 (49%) out of the 253 participants provided their mailing address. Thirteen others (5%) gave their e-mail address, and one (<1%) did not complete the mailing address. The remaining 116 (46%) left the item blank. In the Internet condition, 159 (94%) out of the 170 participants provided their e-mail address; the remaining 11 (6%) left the item blank. Overall, a higher percentage of participants provided their valid address in the Internet condition. The corresponding chi-square analysis was significant, $\chi^2(1) = 92.30$, $p < .001$.

We then investigated how many participants who had been sent the survey returned it. Due to a problem with the surveys that was not detected in time, seven participants in each of the conditions were removed from analyses. Out of the 116 participants in the paper-and-pencil condition who were mailed the survey, 86 (74%) returned it, 25 (22%) did not return it, four (3%) did not have it delivered due to an invalid address, and one (1%) returned it unopened. There was no evidence of random or repeat responding. Of the 152 in the Internet condition that were e-mailed the survey link, 124 (82%) completed the survey, 25 (16%) did not start it, and four (3%) started but did not finish it. Three participants (2%) were removed due to random responding. Two (2%) responded twice; only the first response was included in further analyses. Taken together, the conditions appeared to be equivalent. The chi-square analysis comparing the groups was not statistically significant at $p \leq .025$, $\chi^2(1) = 1.76$, *ns*. Overall, although groups differed for providing valid contact information, they seemed to be equivalent for completing the study.

Missing data. Our examination of missing data included five chi-square analyses; we used an alpha correction of $p \leq .01$ (.05/5). For completion time, four participants (5%) in the paper-and-pencil condition and one (1%) in the Internet condition did not enter a usable date and time. One person (1%) in the Internet condition did not provide a response to comfort completing paper-and-pencil questionnaires. Regarding the four demographic variables (gender, age, ethnic/racial identity, and year in school), three participants (3%) in the Internet condition did not provide their year in school ($n = 2$) or ethnic/racial identity ($n = 1$). Missing data appeared to be equivalent. The chi-square analyses were not statistically significant at $p \leq .01$ for completion time, $\chi^2(1) = 2.97$, *ns*, comfort completing paper-and-pencil questionnaires, $\chi^2(1) = 0.74$, *ns*, or demographic variables, $\chi^2(1) = 2.24$, *ns*.

We next examined missing data for the main variables. Nine (10%) of participants in the paper-and-pencil condition and 23 (20%) in the Internet condition were missing data; the groups appeared to be equivalent. The chi-square analysis was not significant at $p \leq .01$, $\chi^2(1) = 3.16$, *ns*. Out of the 32 participants with missing data, the majority were missing either one ($n = 23$, 72%) or two data points ($n = 4$, 13%), with the remaining participants, all in the Internet condition, missing three ($n = 2$, 6%), six, 10, or 15 points ($n = 1$, 3% each) out of the possible 111 data points per person. Overall, presence and amount of missing data appeared to be equivalent.

Completion time. Means and standard deviations for completion time are found in Table 4. We used mean equivalence testing to compare groups. Frequency analyses yielded three outliers in the paper-and-pencil condition of 1,414 min (approximately 24 hr), 2,615 min (approximately 43½ hr), and 6,823 min (approximately 114 hr), and one outlier in the Internet condition of 157 min (approximately 2½ hr). These outliers were removed for the mean equivalence test. The mean equivalence test indicated that the groups were not equivalent (see Table 5); they reached equivalence at 50%. An ANOVA was then conducted as the appropriate NHST. Levene's test was significant, $F(1, 192) = 34.27$, $p < .001$, and the ANOVA was also significant, $F(1, 192) = 34.27$, $p < .001$, $\eta_p^2 = .17$, with those in the paper-and-pencil condition taking significantly longer to complete the study than those in the Internet condition. These findings indicated that completion time was not equivalent between conditions.

Comfort completing questionnaires. We examined comfort completing questionnaires using paper-and-pencil and the Internet through mean equivalence testing; means and standard deviations for these variables are shown in Table 4. The mean equivalence tests indicated comparability for paper-and-pencil comfort and Internet comfort (see Table 5). For the NHST MANOVA, Levene's test was significant for comfort completing paper-and-pencil questionnaires, $F(1, 200) = 13.82$, $p < .001$. The MANOVA was not significant at $p \leq .017$, $F(2, 199) = 3.90$, *ns*, $\eta_p^2 = .04$. Overall, the conditions were equivalent.

Taken together, our results generally indicated quantitative and qualitative equivalence. They also indicated that two of the aspects of auxiliary equivalence were comparable (i.e., missing data and comfort completing questionnaires). The conditions differed on response rates, with those in the paper-and-pencil condition being less likely to provide a valid address than those in the Internet condition; however, the two conditions were similar in regard to returning the completed questionnaires. There was also a difference in total completion time, with those in the paper-and-pencil condition taking longer than those in the Internet condition.

General Discussion

Summary and Implications of Main Findings

The results of the two present studies generally supported quantitative and qualitative equivalence, as well as some aspects of auxiliary equivalence, between paper-and-pencil and Internet conditions for self-report survey-based measures (i.e., personality, social desirability, computer self-efficacy, demographics, and, in Study 2, personal growth initiative). This implies that self-report survey-based measures can generally be administered through the Internet with good (i.e., equivalent to paper-and-pencil) results.

However, there were also some notable differences regarding the auxiliary equivalence aspects of completion time (Studies 1 and 2), missing data (Study 1), and response rates (Study 2), which need to be taken into account in certain situations.

Quantitative equivalence. The means for the main variables—Extraversion, Neuroticism, Openness to Experience, Agreeableness, Conscientiousness, Social Desirability, CSE Beginner, CSE Conceptual, CSE Mainframe, and, in Study 2, Readiness for Change, Planfulness, Intentional Behavior, and Using Resources—were generally equivalent across paper-and-pencil and Internet conditions, both when completed in lab or take-home settings with experimenter contact (Study 1) and when participants were mailed or e-mailed the survey without experimenter contact (Study 2). Study 1 indicated a few inconclusive findings for CSE Beginner, CSE Mainframe, and Neuroticism in which the mean equivalence test showed differences at the predetermined 20% interval, but the NHST was not significant; the effects were equivalent at intervals ranging from 22%–25%. Almost all of these inconclusive findings involved the lab paper-and-pencil condition.

Qualitative equivalence. Internal consistency analyses indicated all measures were at least in the marginal range (≥.60), with most being adequate (≥.70). These findings are consistent with previous internal consistency estimates for the respective scales (e.g., Barbeite & Weiss, 2004; Barger, 2002; Buchanan, Johnson, et al., 2005; Robitschek et al., 2012). The internal consistency comparisons in both studies indicated equivalence across conditions, with the exception of Conscientiousness and Agreeableness in Study 1 and Using Resources in Study 2. For the intercorrelations, the only significantly different correlation was Neuroticism and Openness in Study 2; this difference was due to direction rather than magnitude.

Auxiliary equivalence. We first examined response rates for the different conditions. In Study 1, almost all participants returned the completed questionnaires, regardless of condition. In Study 2, in which participants did not have personal contact with the researcher, response rates were different for providing an address, as many participants in the paper-and-pencil condition did not provide their physical mailing address. Perhaps participants were more concerned with providing a physical address than an e-mail address due to stronger privacy and security concerns. However, of those receiving the survey, almost all returned it in both conditions.

Examinations of missing data in both studies indicated that completion time, comfort completing questionnaires, and demographic data were reported by almost all of the participants, and there were no differences across conditions. Regarding missing data points for the main variables in Study 1, more participants in both Internet conditions, but particularly those in the lab Internet group, were missing data points than in the paper-and-pencil conditions. However, the actual amount of missing data was low. Study 2 indicated comparability in terms of the number of participants missing data, with only slightly more participants missing data in the Internet condition than in the paper-and-pencil condition, as well as equivalence in terms of number of missing data points. Overall, the amount of missing data was trivial across all studies and conditions, with a trend toward slightly more missing data in the Internet conditions. In most situations, the presence of small amounts of missing data is not a particularly large issue, although paper-and-pencil formats may be more appropriate than the Internet when it is essential for participants to complete all items.

The results for the average time to complete the questionnaires were mostly similar in the two studies. For Study 1, those in the take-home paper-and-pencil condition took several minutes longer to complete the survey than those in the other three conditions. In Study 2, those in the paper-and-pencil condition took more time to complete the survey than those in the Internet condition. Overall, if participants complete paper-and-pencil assessments in a setting of their choice, they take longer to complete the surveys, including compared to completing the surveys in a setting of their choice via the Internet. This might be due to the lack of perceived pressure to finish in a timely fashion or a higher likelihood of becoming distracted and setting aside the survey that is unique to this particular setting, as both the presence of an experimenter and the possibility that a website will "time out," necessitating participants to restart the study, would motivate the participant to not interrupt or postpone the assessment.

Participants also generally indicated high levels of comfort completing questionnaires in paper-and-pencil and Internet conditions. Results of both studies showed mean equivalence across conditions. This indicates that, in general, college students are comfortable completing surveys using both modes, which is consistent with their exposure to technology (see McCoy, 2010).

Limitations and Future Research

There are some limitations in our study which should be addressed in future research. First, a few of our statistical results suggested a difference using equivalence tests that was not significant using NHST, indicating that the findings were inconclusive (Rogers et al., 1993). Consequently, we cannot make judgments about the equivalence of these particular conditions. This has occurred in previous research on paper-and-pencil and Internet conditions (Lewis et al., 2009). Due to the small amount of literature examining paper-and-pencil and Internet equivalence, this is an important area for more specific future research on equivalence testing methodology.

Second, auxiliary equivalence had mixed outcomes. In the current studies, the comparability of the measures was not impacted by auxiliary equivalence. However, the influence of auxiliary equivalence on psychologists will differ depending on what is done. Therefore, research is needed to systematically evaluate auxiliary equivalence, similar to the ways quantitative and qualitative equivalence have been defined and evaluated in the literature (e.g., Buchanan, 2007; Meyerson & Tryon, 2003).

Third, participants were all college students. The results may not generalize to other populations, particularly ones that have less access to, experience with, and knowledge of Internet-related technology or are just generally less comfortable with using it. For example, participants with little experience using computers may have a lower response rate when in the Internet conditions compared to the paper-and-pencil conditions. Also, our participants received credit for participating in the study, so the results might not generalize to unpaid surveys and unsolicited survey requests common both in the form of "junk" mail and on many websites. Thus, future research should seek to replicate our results with different populations and compensation schedules.

Finally, our studies were based on short, self-report survey-based measures. Although these were chosen based on the prominence of such measures in social science (e.g., Berends, 2006) and

Internet data collections (e.g., Buchanan, 2007), they do not reflect many of the types of research in which psychologists are engaged, for example, studies involving reaction times, attention, or mental health treatments. Literature concerning the comparability of non-survey-based types of data collection across in-lab and Internet conditions (e.g., McGraw, Tew, & Williams, 2000; Preckel & Thiemann, 2003; Steele et al., 2009; see Buchanan, 2007) is relatively scarce. Thus, further research is warranted in this area.

Recommendations and Conclusion

Based on the results of our studies, we are cautiously optimistic regarding the general equivalence of paper-and-pencil and Internet self-report survey-based questionnaires. More often than not, psychologists can likely administer a measure or assessment that has been validated for one mode in the other mode without much adverse impact, even if the setting of administration is also different. However, when given the choice between administering a measure that is validated for that particular mode and one that is not, we recommend using the former, with the caveat that psychologists should also take into account the purpose of the assessment and any related setting conditions before making a final determination. For example, if it is essential for those completing the measure to do so in one sitting, an Internet format might be more likely to yield usable results than mailed paper-and-pencil; if it is critical that every item is answered, then the opposite is true.

Overall, the current research has shown the equivalence of paper-and-pencil and Internet data collection methods in two studies with different settings and with and without researcher contact, except for some aspects of auxiliary equivalence. The differences found when assessing auxiliary equivalence did not affect the overall quantitative and qualitative equivalence of the measures. Although research continues to be needed in this area, we expect that most self-report survey-based psychological measures can successfully transition from the traditional paper-and-pencil format to the Internet.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43,* 335–344. doi:10.1016/j.jrp.2008.12.013

Bandilla, W., Bosnjak, M., & Altdorfer, P. (2003). Survey administration effects? A comparison of Web-based and traditional written self-administered surveys using the ISSP environment module. *Social Science Computer Review, 21,* 235–243. doi:10.1177/0894439303021002009

Barak, A., & Cohen, L. (2002). Empirical investigation of an online version of the Self-Directed Search. *Journal of Career Assessment, 10,* 387–400. doi:10.1177/1069072702238402

Barbeite, F. G., & Weiss, E. M. (2004). Computer self-efficacy and anxiety scales for an Internet sample: Testing measurement equivalence of existing measures and development of new scales. *Computers in Human Behavior, 20,* 1–15. doi:10.1016/S0747-5632(03)00049-9

Barger, S. D. (2002). The Marlowe-Crowne affair: Short forms, psychometric structure, and social desirability. *Journal of Personality Assessment, 79,* 286–305. doi:10.1207/S15327752JPA7902_11

Berends, M. (2006). Survey methods in educational research. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 623–640). Mahwah, NJ: Erlbaum.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review, 78,* 161–188. doi:10.1111/j.1751-5823.2010.00112.x

Boyer, K. K., Olson, J. R., Calantone, R. J., & Jackson, E. C. (2002). Print versus electronic surveys: A comparison of two data collection methodologies. *Journal of Operations Management, 20,* 357–373. doi:10.1016/S0272-6963(02)00004-9

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice, 33,* 148–154. doi:10.1037/0735-7028.33.2.148

Buchanan, T. (2003). Internet-based questionnaire assessment: Appropriate use in clinical contexts. *Cognitive Behaviour Therapy, 32,* 100–109. doi:10.1080165060703100000957

Buchanan, T. (2007). Personality testing on the Internet: What we know, and what we do not. In A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of Internet psychology* (pp. 447–459). Oxford, England: Oxford University Press.

Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., & Scholey, A. B. (2005). Nonequivalence of on-line and paper-and-pencil psychological tests: The case of the prospective memory questionnaire. *Behavior Research Methods, 37,* 148–154. doi:10.3758/BF03206409

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment, 21,* 115–127. doi:10.1027/1015-5759.21.2.115

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology, 90,* 125–144. doi:10.1348/000712699161189

Cantrell, M. A., & Lupinacci, P. (2007). Methodological issues in online data collection. *Journal of Advanced Nursing, 60,* 544–549. doi:10.1111/j.1365-2648.2007.04448.x

Carini, R. M., Hayek, J. C., Kuh, G. D., Kennedy, J. M., & Ouimet, J. A. (2003). College student responses to Web and paper surveys: Does mode matter? *Research in Higher Education, 44,* 1–19. doi:10.1023/A:1021363527731

Church, A. H. (2001). Is there a method to our madness? The impact of data collection methodology on organizational survey results. *Personnel Psychology, 54,* 937–969. doi:10.1111/j.1744-6570.2001.tb00238.x

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi:10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p <$. 05). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

Cole, M. S., Bedeian, A. G., & Feild, H. S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods, 9,* 339–368. doi:10.1177/1094428106287434

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly, 19,* 189–211. doi:10.2307/249688

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI– qR) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Odessa, FL: Psychological Assessment Resources.

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60,* 1–10. doi:10.1002/jclp.10217

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments & Computers, 34,* 177–180. doi:10.3758/BF03195440

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments & Computers, 31,* 572–577. doi:10.3758/BF03200737

Diener, E. (2009). Positive psychology: Past, present, and future. In S. J. Lopez & C. R. Snyder (Eds.), *Oxford handbook of positive psychology* (pp. 7–11). New York, NY: Oxford University Press.

Dolnicar, S., Laesser, C., & Matus, K. (2009). Online versus paper: Format effects in tourism surveys. *Journal of Travel Research, 47,* 295–316. doi:10.1177/0047287508326506

Drummond, R. J., & Jones, K. D. (2010). *Assessment procedures for counselors and helping professionals* (7th ed.). Boston, MA: Pearson.

Epstein, J., Klinkenberg, W. D., Wiley, D., & McKinley, L. (2001). Insuring sample equivalence across Internet and paper-and-pencil assessments. *Computers in Human Behavior, 17,* 339–346. doi:10.1016/S0747-5632(01)00002-4

Evans, D. C., Garcia, D. J., Garcia, D. M., & Baron, R. S. (2003). In the privacy of their own homes: Using the Internet to assess racial bias. *Personality and Social Psychology Bulletin, 29,* 273–284. doi:10.1177/0146167202239052

Fischer, D. G., & Fick, C. (1993). Measuring social desirability: Short forms of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 53,* 417–424. doi:10.1177/0013164493053002011

Fouladi, R. T., McCarthy, C. J., & Moller, N. P. (2002). Paper-and-pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. *Assessment, 9,* 204–215. doi:10.1177/10791102009002011

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance-testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 149–169). New York, NY: Guilford Press.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe,* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84–96. doi:10.1016/j.jrp.2005.08.007

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59,* 93–104. doi:10.1037/0003-066X.59.2.93

Harrison, A. W., & Rainer, R. K. (1992). An examination of the factor structures and concurrent validities for the Computer Attitude Scale, the Computer Anxiety Rating Scale, and the Computer Self-Efficacy Scale. *Educational and Psychological Measurement, 52,* 735–745. doi:10.1177/0013164492052003024

Heath, N. M., Lawyer, S. R., & Rasmussen, E. B. (2007). Web-based versus paper-and-pencil course evaluations. *Teaching of Psychology, 34,* 259–261. doi:10.1080/00986280701700433

Herrero, J., & Meneses, J. (2006). Short Web-based versions of the Perceived Stress (PSS) and Center for Epidemiological Studies-Depression (CES-D) Scales: A comparison to pencil and paper responses among Internet users. *Computers in Human Behavior, 22,* 830–846. doi:10.1016/j.chb.2004.03.007

Hewson, C., & Charlton, J. P. (2005). Measuring health beliefs on the Internet: A comparison of paper and Internet administrations of the Multidimensional Health Locus of Control Scale. *Behavior Research Methods, 37,* 691–702. doi:10.3758/BF03192742

Huang, H.-M. (2006). Do print and web surveys provide the same results? *Computers in Human Behavior, 22,* 334–350. doi:10.1016/j.chb.2004.09.012

International Personality Item Pool. (2011). *A scientific collaboratory for the development of advanced measures of personality and other individual differences*. Retrieved from http://ipip.ori.org/

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments & Computers, 31,* 433–438. doi:10.3758/BF03200723

Khorrami-Arani, O. (2001). Researching computer self-efficacy. *International Education Journal, 2,* 17–25.

Kiernan, N. E., Kiernan, M., Oyler, M. A., & Gilles, C. (2005). Is a web survey as effective as a mail survey? A field experiment among computer users. *American Journal of Evaluation, 26,* 245–252. doi:10.1177/1098214005275826

Kramer, G. P., Bernstein, D. A., & Phares, V. (2009). *Introduction to clinical psychology* (7th ed.). Upper Saddle River, NJ: Pearson.

Lewis, I. M., Watson, B. C., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology, 61,* 107–116. doi:10.1080/00049530802105865

McCabe, S. E. (2004). Comparison of web and mail surveys in collecting illicit drug use data: A randomized experiment. *Journal of Drug Education, 34,* 61–72. doi:10.2190/4HEY-VWXL-DVR3-HAKV

McCoy, C. (2010). Perceived self-efficacy and technology proficiency in undergraduate college students. *Computers & Education, 55,* 1614–1617. doi:10.1016/j.compedu.2010.07.003

McDonald, H., & Adam, S. (2003). A comparison of online and postal data collection methods in marketing research. *Marketing Intelligence & Planning, 21,* 85–95. doi:10.1108/02634500310465399

McGrath, R. E. (2011). *Quantitative models in psychology*. Washington, DC: American Psychological Association. doi:10.1037/12316-000

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science, 11,* 502–506. doi:10.1111/1467-9280.00296

Meyerson, P., & Tryon, W. W. (2003). Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments & Computers, 35,* 614–620. doi:10.3758/BF03195541

Miyazaki, A. D., & Taylor, K. A. (2008). Researcher interaction biases and business ethics research: Respondent reactions to researcher characteristics. *Journal of Business Ethics, 81,* 779–795. doi:10.1007/s10551-007-9547-5

Murphy, C. A., Coover, D., & Owen, S. V. (1989). Development and validation of the Computer Self-Efficacy Scale. *Educational and Psychological Measurement, 49,* 893–899. doi:10.1177/001316448904900412

Naus, M. J., Philipp, L. M., & Samsi, M. (2009). From paper to pixels: A comparison of paper and computer formats in psychological assessment. *Computers in Human Behavior, 25,* 1–7. doi:10.1016/j.chb.2008.05.012

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301. doi:I0.1037//1082-989X.S.2.241

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological research on the Internet. *Journal of Social Issues, 58,* 161–176. doi:10.1111/1540-4560.00254

Patterson, S. D., & Jones, B. (2006). Bioequivalence: A review of study design and statistical analysis for orally administered products. *International Journal of Pharmaceutical Medicine, 20,* 243–250. doi:10.2165/00124363-200620040-00004

Pettit, F. A. (2002). A comparison of World-Wide Web and paper-and-pencil personality questionnaires. *Behavior Research Methods, Instruments & Computers, 34,* 50–54. doi:10.3758/BF03195423

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology, 56,* 733–752. doi:10.1111/j.1744-6570.2003.tb00757.x

Preckel, F., & Thiemann, H. (2003). Online- versus paper-pencil-version of a high potential intelligence test. *Swiss Journal of Psychology, 62,* 131–138. doi:10.1024//1421-0185.62.2.131

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology, 38,* 119–125. doi:10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118>3.0.CO;2-I

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84,* 754–775. doi:10.1037/0021-9010.84.5.754

Robitschek, C., Ashton, M. W., Spering, C. C., Geiger, N., Byers, D., Schotts, C., & Thoen, M. A. (2012). Development and psychometric evaluation of the Personal Growth Initiative Scale–II. *Journal of Counseling Psychology, 59,* 274–287. doi:10.1037/a0027310

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565. doi:10.1037/0033-2909.113.3.553

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York, NY: Appleton-Century-Crofts.

Rusticus, S. A., & Lovato, C. Y. (2011). Applying tests of equivalence for multiple group comparisons: Demonstration of the confidence interval approach. *Practical Assessment, Research & Evaluation, 16,* 1–6.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657–680.

Schulenberg, S. E., & Yutrzenka, B. A. (1999). The equivalence of computerized and paper-and-pencil psychological instruments: Implications for measures of negative affect. *Behavior Research Methods, Instruments & Computers, 31,* 315–321. doi:10.3758/BF03207726

Schulenberg, S. E., & Yutrzenka, B. A. (2001). Equivalence of computerized and conventional versions of the Beck Depression Inventory–II (BDI-II). *Current Psychology, 20,* 216–230. doi:10.1007/s12144-001-1008-1

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3,* 403–411. doi:10.1037/1082-989X.3.4.403

Seligman, M. E. P., Steen, T. A., Park, N., & Peterson, C. (2005). Positive psychology progress: Empirical validation of interventions. *American Psychologist, 60,* 410–421. doi:10.1037/0003-066X.60.5.410

Sethuraman, R., Kerin, R. A., & Cron, W. L. (2005). A field study comparing online and offline data collection methods for identifying product attribute preferences using conjoint analysis. *Journal of Business Research, 58,* 602–610. doi:10.1016/j.jbusres.2003.09.009

Shih, T.-H., & Fan, X. (2008). Comparing response rates from Web and mail surveys: A meta-analysis. *Field Methods, 20,* 249–271. doi:10.1177/1525822X08317085

Smith, B., Smith, T. C., Gray, G. C., & Ryan, M. A. K. (2007). When epidemiology meets the Internet: Web-based surveys in the Millennium Cohort Study. *American Journal of Epidemiology, 166,* 1345–1354. doi:10.1093/aje/kwm212

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments & Computers, 29,* 496–505. doi:10.3758/BF03210601

Smither, J. W., Walker, A. G., & Yap, M. K. T. (2004). An examination of the equivalence of Web-based versus paper-and-pencil upward feedback ratings: Rater- and ratee-level analyses. *Educational and Psychological Measurement, 64,* 40–61. doi:10.1177/0013164403258429

Stanton, J. M. (1998). An empirical assessment of data collection using the Internet. *Personnel Psychology, 51,* 709–725. doi:10.1111/j.1744-6570.1998.tb00259.x

Steele, R. M., Mummery, W. K., & Dwyer, T. (2009). A comparison of face-to-face or Internet-delivered physical activity intervention on targeted determinants. *Health Education & Behavior, 36,* 1051–1064. doi:10.1177/1090198109335802

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning, 19,* 193–198. doi: http://dx.doi.org/10.1016/0149-7189(96)00011-0

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371–386. doi:10.1037/1082-989X.6.4.371

Tryon, W. W., & Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001). reduction factor. *Psychological Methods, 13,* 272–277. doi:10.1037/a0013158

Vallejo, M. A., Mañanes, G., Comeche, M. A., & Díaz, M. I. (2008). Comparison between administration via Internet and paper-and-pencil administration of two clinical instruments: SCL-90-R and GHQ-28. *Journal of Behavior Therapy and Experimental Psychiatry, 39,* 201–208. doi:10.1016/j.jbtep.2007.04.001

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44,* 315–327. doi:10.1016/j.jrp.2010.03.003

Van de Vijver, F. J. R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79,* 852–859. doi:10.1037/0021-9010.79.6.852

Voiskounsky, A. E. (2008). Flow experience in cyberspace: Current studies and perspectives. In A. Barak (Ed.), *Psychological aspects of cyberspace: Theory, research, applications* (pp. 70–101). New York, NY: Cambridge University Press.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32,* 741–744. doi:10.2307/2529259

Whitaker, B. G. (2007). Internet-based attitude assessment: Does gender affect measurement equivalence? *Computers in Human Behavior, 23,* 1183–1194. doi:10.1016/j.chb.2004.11.016

Whittier, D. K., Seeley, S., & St. Lawrence, J. S. (2004). A comparison of Web- with paper-based surveys of gay and bisexual men who vacationed in a gay resort community. *AIDS Education and Prevention, 16,* 476–485. doi:10.1521/aeap.16.5.476.48735

Wiechmann, D., & Ryan, A. M. (2003). Reactions to computerized testing in selection contexts. *International Journal of Selection and Assessment, 11,* 215–229. doi:10.1111/1468-2389.00245

Wood, E., Nosko, A., Desmarais, S., Ross, C., & Irvine, C. (2006). Online and traditional paper-and-pencil survey administration: Examining experimenter presence, sensitive material and long surveys. *Canadian Journal of Human Sexuality, 15,* 147–155.

Zhang, Y. (2000). Using the Internet for survey research: A case study. *Journal of the American Society for Information Science, 51,* 57–68. doi:10.1002/(SICI)1097-4571(2000)51:1<57::AID-ASI9>3.0.CO;2-W

Zook, A., & Sipps, G. J. (1985). Cross-validation of a short form of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology, 41,* 236–238. doi:10.1002/1097-4679(198503)41:2<236::AID-JCLP2270410217>3.0.CO;2-H

Received January 17, 2012
Revision received November 15, 2012
Accepted November 21, 2012 ■