Socially Desirable Responding in Web-Based Questionnaires: A Meta-Analytic Review of the Candor Hypothesis

Timo Gnambs

Leibniz Institute for Educational Trajectories

Kai Kaspar

University of Cologne

Word count: 10,956

Author Note

Timo Gnambs, Leibniz Institute for Educational Trajectories, Germany; Kai Kaspar, Department of Psychology, University of Cologne, Germany.

We are grateful to Jennifer Lindzus for her aid during the literature search.

Correspondence concerning this article should be addressed to Timo Gnambs, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, Phone: +49 (0)951 / 8633420; Email: timo.gnambs@lifbi.de

Accepted for publication in *Assessment*.

Abstract

Unproctored, web-based assessments supposedly reduce social desirability distortions in self-report questionnaires because of an increased sense of privacy amongst participants. Three random-effects meta-analyses focusing either on social desirability (*k* = 30, total *N* = 3,746), the Big Five of personality (*k* = 66, total *N* = 2,951), or psychopathology (*k* = 96, total *N* = 16,034) compared social desirability distortions of self-reports across computerized and paper-and-pencil administration modes. Overall, a near-zero effect, Δ = 0.01, was obtained that did not indicate less socially desirable responding in computerized assessments. Moreover, moderator analyses did not identify differential effects for proctored and unproctored procedures. Thus, paper-and-pencil and computerized administrations of self-report scales yield comparable mean scores. Unproctored web-based surveys do not offer an advantage with regard to socially desirable responding in self-report questionnaires.

Keywords: social desirability; survey mode; web-based; response bias; meta-analysis

Socially Desirable Responding in Web-Based Questionnaires: A Meta-Analytic Review of the Candor Hypothesis

The popularity of computerized devices in everyday life has facilitated a variety of new assessment procedures in psychological research and practice (cf. Gosling & Mason, 2015; Tippins, 2015; Trull & Ebner, 2013; Vinciarelli & Mohammadi, 2014). Particularly the use of web-based surveying and testing (WBT)—that is, unproctored computerized tests administered over the Internet—has risen continuously during recent decades (Couper, 2011). In view of these technological changes, concerns have been voiced about whether responses from computerized tests are comparable to those from traditional procedures. The "candor" hypothesis (Buchanan, 2000, 2001) postulates that WBT should result in less socially distorted responses because of an increased sense of privacy provided by computerized environments. However, existing experimental studies provided mixed results in this respect. Some studies identified less socially desirable distortions in WBT (Kays, Gathercoal, & Burhow, 2012); others reported no (Weigold, Weigold, & Russell, 2013) or even the opposite effects (Vecchione, Alessandri, & Barbaranelli, 2012). Therefore, this review summarizes the effects of computerized surveys, particularly in unproctored web-based settings, on socially desirable responding. Three meta-analyses examine mean-level differences between computerized and paper-and-pencil tests of social desirability (Meta-Analysis I), the Big Five of personality (Meta-Analysis II), and psychopathological symptoms (Meta-Analysis III).

Dimensions of Psychological Survey Modes

Survey modes can vary along several dimensions, such as the degree of interviewer involvement, the adopted survey technology, and the privacy afforded to respondents (cf. Couper, 2011; Groves et al., 2009). WBT as a variant of computerized tests are characterized by assessment situations without the presence of a human supervisor. Rather, the assessment conditions are unstandardized and remain the responsibility of the respondent. Thus, unproctored WBT typically exhibits larger variations in test-taking conditions such as the assessment situation (e.g., at home or at work) or environmental distractions (e.g., noise). However, for economic reasons WBT has become the de facto standard in many fields of research. For example, commercial market research companies administer up to seven times as many surveys over the Internet as by mail (ADM, 2015). Similarly, in 2009 and 2010 about 11% of all empirical articles published in major social psychological journals included at least one web-based sample (Skitka, Sargis, & McKeeveer, 2013). However, the last decade also registered a sharp increase in mixed-mode designs that assign respondents to different survey modes (De Leeuw & Hox, 2011). For example, a web-based study might be complemented by a postal survey to reach respondents with no or limited Internet access. Hence, the question arises whether survey mode-specific conditions systematically affect respondents' answers.

Mode Effects and Social Desirability

Although WBT offers various advantages including, inter alia, access to a large number of hard-to-reach participants (e.g., individuals with rare psychological disorders), they provide few benefits in terms of improved psychometric properties. Paper-and-pencil and web-based tests typically show similar factor structures and reliabilities (e.g., Bjorner et al., 2014; Swahney & Cigularov, 2014; Vecchione et al., 2012; Weigold et al., 2013); even the predictive validities do not appear significantly different (Beaty et al., 2011). In contrast, the results with regard to mean-level equivalence are not entirely consistent: Some authors concluded that the presentation mode does not affect latent (Chuah, Drasgow, & Roberts, 2006) or score means (Weigold et al., 2013); others observed slightly different means in web-based as compared to paper-and-pencil tests (e.g., Aluja, Rossier, & Zuckerman, 2007; Ployhart, Weekley, Holtz, & Kemp, 2003). These differences are typically interpreted as resulting from specific response styles because unproctored WBT is assumed to enhance people's readiness to engage in less socially desirable responding (Buchanan, 2000, 2001).

Previous research suggested two central factors explaining a social desirability bias: On the one hand, social desirability might be a consequence of stable individual differences in the need for social approval (Crowne & Marlowe, 1960) or the honesty-humility trait (de Vries, Zettler, & Hilbig, 2014) and thus the disposition for impression management. On the other hand, an individual's propensity to disclose personally sensitive information might also be determined by transient situational characteristics (John, Acquisti, & Loewenstein, 2011). For example, people tend to disclose more unfavorable information about themselves to others if legal conditions facilitate an honest response (Galletly & Pinkerton, 2006) or if test settings are perceived as lending high levels of privacy (Joinson & Paine, 2006). As a consequence, even supposedly unrelated cues in the assessment procedure might increase self-reports of potentially harmful content. For example, with the advent of computerized testing researchers were hoping for a reduction in socially desirable responding in self-reports (Fox & Schwartz, 2002), because elimination of the interviewer was supposed to reduce perceived social pressure and to increase the feeling of anonymity among respondents. However, initial meta-analyses (Dwight & Feigelson, 2000; Richman, Kiesler, Weisband & Drasgow, 1999) did not find a direct link between the presentation mode (computer vs. paper) and social desirability effects. Thus, a simple switch from paper-and-pencil to computerized survey modes does not necessarily reduce social desirability distortions. Rather, context factors seem to moderate this effect. The more a respondent feels that privacy, anonymity, and data security are assured, the more he or she is likely to provide personal sensitive information. For example, Joinson, Woodley, and Paine (2007) revealed a decreased willingness to divulge one's income—an item typically seen as rather sensitive—when respondents had to enter a username and password before getting access to a web-based survey as compared to users receiving anonymous links to the questionnaire. Also, participants spontaneously reported more personal information in web-based discussion groups—particularly when they were also visually anonymous (thus, there was no video transmission)—than in comparable face-to-face groups (Joinson, 2001). Similar results were obtained using validated self-report scales (Fox & Schwartz, 2002; Joinson, 1999). Moreover, the more the social presence of the interviewer is reduced, the greater the truthfulness of respondents becomes, because peripheral cues such as the interviewer's sex do not affect responses in this setting (Tourangeau & Yan, 2007). Consequently, computerized testing per se does not necessarily reduce tendencies towards socially desirable behavior. Rather, the unproctored nature of WBT might be the key feature.

To this effect, the "candor" hypothesis (Buchanan, 2000, 2001) postulates that WBT leads to less socially distorted responses because the assessment situation is perceived as more anonymous and less judgmental. Indeed, compared to traditional modes, respondents tend to report higher levels of alcohol consumption, more illicit drug use, and more frequent sexual activities in WBT (Källmén, Sinadinovic, Berman, & Wennberg, 2011; Kays et al., 2012). However, experimental studies comparing the degree of impression management or studying social desirability effects inferred from other self-reports could not unequivocally confirm this effect. Some authors observed increased self-disclosure when surveys were administered over the Internet, whereas others found only small or even null effects (e.g., Carlbring et al., 2007; Fogarty, Jonas, & Parker, 2013; Risko, Quilty, & Oakman, 2006). For example, students evaluate instructors and their courses more critically in web-based as compared to paper-and-pencil questionnaires (Fogarty et al., 2013), and people also report slightly higher levels of depression on the Internet (Carlbring et al., 2007), whereas Risko and colleagues (2006) found no evidence of such mode differences for various measures of social desirability. Therefore, we propose two hypotheses that are examined in three meta-analyses:

H1: Self-reported mean scores of socially undesirable traits are higher in computerized as compared to paper-and-pencil tests.

H2: The difference in self-reported mean scores of socially undesirable traits between computerized and paper-and-pencil tests is larger for unproctored assessments as compared to proctored ones.

Overview of Meta-Analyses

Three meta-analyses of mode experiments examined response distortions in WBT. Computerized, particularly unproctored WBT was expected to result in less socially desirable responding than comparable paper-and-pencil tests. Meta-Analysis I focused on social desirability scales for the analysis of cross-mode differences. The other studies adopted an indirect approach and inferred social desirability effects from instruments measuring the Big Five of personality (Meta-Analysis II) and psychopathological symptoms (Meta-Analysis III). Following prevalent conventions in the interpretation of effect sizes (Ferguson, 2009), standardized mean differences of at least *d* = .41 are viewed as indicative of practically relevant differences.

Meta-Analysis I: Explicit Measures of Social Desirability

Method

Data source. Eligible studies on socially desirable responding in paper-and-pencil and computerized assessments were identified by searching bibliographic databases (PsycINFO, Psyndex, Psychology & Behavioral Sciences Collection, EconLit, Business Source Complete, ProQuest Dissertations & Theses Database) using the keywords *social desirability*, *self-disclosure*, *impression management*, or *response distortion* in combination with *computer-based*, *computerized*, *web-based*, or *internet-based* and *paper, mail,* or *postal*. Additional studies were located from a comparable search in Google Scholar. For the latter, we examined all 1,000 results that are returned by the search engine (Boeker, Vach, & Motschall, 2013). The entire search process is summarized in Table S1 of the online supplement.

Inclusion criteria. Studies were included in the meta-analysis according to five criteria: (a) A validated social desirability scale was administered. (b) The study adopted an experimental design that either randomly administered a paper-and-pencil or computerized version of the instrument in a proctored or an unproctored setting, or provided measures for both modes in a within-subject design. Studies where participants themselves chose their preferred mode of administration were not included. (c) Computer and paper-and-pencil conditions adopted identical administration settings. Studies that administered one test version in an unproctored setting and the other version as a proctored test were not included. Otherwise, moderation analyses regarding the administration setting would not be feasible. (d) The study reported relevant effect sizes or respective information to compute an effect size. Finally, (e) only studies published no earlier than the year 2000 were retained. Although some researchers started experiments on the Internet in the mid-nineties (cf. Bartram, 2000; Musch & Reips, 2000), web-based methods have only gradually gained broader acceptance in psychological research during the last decade (Gosling, Vazire, Srivastava, & John, 2004). To guard against any potential distortions resulting from the unconventional research environment in early studies, our analyses are limited to studies of the current millennium. This has resulted in 12 studies that were eligible for the meta-analysis.

Coding process. From each study, we extracted the sample size and the sample statistics of the social desirability scales to calculate the effect sizes (*M* and *SD*). For studies not reporting the appropriate sample statistics, related information such as correlations, percentages, or test statistics (e.g., *t*-values) were recorded. Based on previous results (Paulhus, 1991; Uziel, 2010), we coded all measures as operationalizing either the impression management or self-deceptive enhancement aspect of social desirability. Measures for impression management included the impression management subscale of the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984), the lie scale of the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1989), the Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960), and the social desirability scale of the Occupational Personality Questionnaire (Saville, Holdsworth, Nyfield, Cramp, & Mabey, 1996). Measures for self-deceptive enhancement included the self-deceptive enhancement subscale of the BIDR and the defensiveness scale of the MMPI. In addition, we extracted several moderators from the primary studies (see Table 1). Because the focal hypothesis referred to unproctored WBT, we also noted whether the test procedure was unsupervised and respondents thus chose their own time and place to take the test. Moreover, we recorded six additional variables and included them as covariates in the analyses:[1] publication year, country of origin, mean age (in years) and percentage of female participants in the sample, sample type (students, patients with psychological disorders, job applicants, or a general community sample), and research design (i.e., within- or between-subject design). All studies were coded by the first author. A random sample of four studies was also independently coded by the second author. The intraclass-correlations (ICC) and Cohen's kappa (κ) for the two codings were rather high and ranged between .99 and 1.00 (*Mdn* = 1.00). Disagreements between the coders were resolved by discussion.

Effect size. The unbiased standardized mean difference *g* was selected as the effect size for the meta-analysis (Hedges, 1981). The effect sizes were calculated in a way that negative effect sizes indicate less social desirability distortion on the computer and positive effect sizes result for increased social desirability on the computer. For studies not reporting appropriate sample statistics to calculate *g,* transformation formulas were applied to derive *g* from percentages (Chinn, 2000) or *t* values (Morris & DeShon, 2002).

Outliers. We identified outliers by using the studentized deleted residual (Viechtbauer & Cheung, 2010). Using an α of 1%, these analyses identified one extreme effect size (about 3% of all included effects). Following prevalent practice (cf. Gnambs, 2014), the impact of the outlier on the pooled effect was reduced by truncating the respective effect size to the bound of the 90% credibility interval of the true effect calculated with a data set from which the outlier had been removed.

Missing values. Six samples did not report the mean age of participants; one sample neglected to specify the percentage of female participants. For these studies, the missing values of moderators were imputed using the median value of the remaining studies.

Meta-analytic procedure. In order to cope with dependencies between effects that resulted from studies reporting multiple mode comparisons (e.g., obtained with different instruments), the random-effects meta-analysis was formulated as a multilevel model where individual effects are nested within studies (Cheung, 2014). To account for sampling error, each effect size was weighted by the inverse of its variance. Inverse variance weights are superior to other weighting schemes and result in more precise estimates of the mean effect (Brannick, Yang, & Cafri, 2011). Studies with extraordinarily large samples (i.e., extreme outliers according to Tukey, 1977) were truncated to the maximum sample size of the remaining studies before calculating the variances of the effect sizes (cf. Gnambs, 2013). Otherwise *g* would have predominantly reflected these large sample studies, giving hardly any weight to the other studies. Corrections for attenuation due to measurement error were not applied because the study focused on the operational equivalence of the scales and how the administration medium affected the observed mean scale scores of the respondents, rather than their theoretical standing on the latent construct.

Heterogeneity in the observed effect sizes was quantified by I^2, indicating the percentage of the total variance in observed effects due to random variance (Higgins, Thompson, Deeks, & Altman, 2003). Following prevalent rules of thumb, I^2 of .25, .50, and .75 indicate low, medium, and high heterogeneity, respectively. In addition, the homogeneity of effects was tested using the *Q* statistic (Cochran, 1954). Because the latter frequently exhibits rather poor power (e.g., Sánchez-Meca & Marín-Martínez, 1997), we relied more on I^2 and whether moderators reduced the random variance. Moderator effects were examined using weighted mixed-effects regression analyses.

Publication bias. We studied the effect of a potential publication bias on the results in two ways. First, meta-regression analyses examined differences in the pooled effects derived from published (i.e., journal articles and books) and unpublished sources (i.e. conference proceedings and theses). Significant differences would indicate that the published research literature was distorted due to the systematic suppression of (most likely small) effects. Second, the funnel plot of the effects sizes was tested for asymmetry using a rank correlation test (Begg & Mazumdar, 1994) and a regression test (Egger, Davey Smith, Schneider, & Minder, 1997). Significant negative effects would indicate systematically missing studies that might have distorted the pooled effect.

Statistical software. All meta-analytic models were estimated with the *metaSEM* software version 0.9.4 (Cheung, 2015). Additional analyses were conducted in *R* version 3.2.0 (R Core Team, 2015).

Results

Sample characteristics. The meta-analysis included 17 independent samples with a total of 3,746 participants reporting 30 effect sizes. Most studies were published in peer-reviewed journals; one study each was reported in a book and a thesis. Approximately 62% of the participants were female. The reported mean age of the samples ranged from 17 to 34 years (*M* = 22.42, *SD* = 5.70). The majority of participants were from the United States (82%) and were classified as students (88%)—only one study included an adult job applicant sample. With regard to the administered instruments, the BIDR contributed about 61% of all effect sizes and the MCSDS about 29%.

Pooled effect. The pooled adjusted effect of computerized assessments on socially desirable responding was $\Delta = 0.03$, $p = .45$ (Table 2) and thus identified no mode effect. The administration mode had no differential effect on impression management, $\Delta = 0.02$, $p = .65$, or self-deceptive enhancement, $\Delta = 0.05$, $p = .45$ (Figure 1). Thus, computerized assessments did not reduce socially desirable responding. Also, meta-regression analyses showed no moderation effect of the administration mode (coded -1 for proctored and 1 for unproctored settings), neither for impression management, $\gamma = 0.03$, $SE = 0.05$, $p = .59$, nor for self-deceptive enhancement, $\gamma = 0.03$, $SE = 0.08$, $p = .73$. Thus, unproctored WBT did not result in less socially desirable responding as compared to proctored computerized assessments.

Sensitivity analyses. The robustness of the results was investigated in several ways. First, one sample simulating a selection process ($\Delta = 0.29$, $p = .19$) was excluded from the analyses. However, with a near null effect, $\Delta = 0.02$, $p = .57$, a meta-analysis on the remaining samples corroborated the previously reported result. Moreover, neither the homogeneity test, $Q = 22.82$, $df = 27$, $p = .69$, nor the I^2 statistic of .03 indicated relevant random variance for the social desirability scales. Thus, it is unlikely that hidden moderators distorted the pooled effect.

Nevertheless, we examined the impact of six[2] between-sample characteristics on potential administration mode differences: the publication year (recoded as deviation from 2014), the research design (coded 1 for within-subject and -1 for between-subject), the origin of the participants (coded 1 for US and -1 for non-US), the percentage of female participants (recoded as deviation from 50), the mean age of the respondents (recoded as deviation from 20), and the administration setting (coded -1 for proctored and 1 for unproctored settings). The coding scheme was adopted to interpret the intercept in the mixed-effects regression model in terms of the mean population effect after controlling for the specified cross-study differences. Moreover, the continuous moderators were recoded in such a way that the intercept reflected the true mode effect in the year 2014 for a sample with a mean age of 20 years and a balanced sex ratio. After accounting for the moderators (Table 3), the intercept and thus the adjusted population effect remained nonsignificant, $\Delta = 0.01$, $p = .92$. The meta-regression analysis identified a single moderating effect for the participants' country of origin: US samples, $\Delta_{\text{predicted}} = 0.10$, showed significantly larger mode effects than non-US samples, $\Delta_{\text{predicted}} = -0.08$. However, neither of the predicted effect sizes reached a practically relevant magnitude (Ferguson, 2009). Thus, the null finding from the previous section was replicated after controlling for several between-sample differences.

Finally, we also examined whether the type of the administered social desirability scale might have affected the reported results. For that purpose, we created two dummy-coded variables that indicated either the administration of the BIDR or the MCSDS. However, a corresponding meta-regression analysis did not identify any different administration mode effects for the BIDR, γ = -0.02, *SE* = 0.08, *p* = .83, or the MCSDC, γ = 0.07, *SE* = 0.10, *p* = .47. Neither the BIDR, Δ = 0.00, *p* = .95, nor the MCSDS, Δ = 0.09, *p* = .25, showed less socially desirable responding in computerized assessments.

Publication bias. In order to investigate whether there was a potential publication bias, effect sizes extracted from published sources were compared to effects from unpublished sources. However, similar effects emerged for published, Δ = 0.02, *p* = .62, and unpublished effect sizes, Δ = 0.08, *p* = .41. Moreover, a visual inspection of the funnel plot (left plot in Figure 2) did not indicate any publication bias, but revealed a largely symmetrical distribution around the population effect. Finally, the non-significant rank correlation, τ = -.16, *p* = .22, and regression tests, *B* = -0.05, *SE* = 0.61, *p* = .94, for funnel plot asymmetry corroborated the lack of evidence regarding a potential publication bias.

Meta-Analysis II: Big Five

Method

In contrast to Meta-Analysis I, the second meta-analysis used indirect indicators of social desirability effects by means of the Big Five of personality (conscientiousness, agreeableness, emotional stability, openness to experiences, and extraversion). In line with Kuncel and Tellegen (2009; see also Paunonen & LeBel, 2012) socially desirable responding was viewed as a deliberate overreporting of favorable characteristics; hence, higher mean levels of the positively evaluated sides of the five traits are assumed to indicate stronger social desirability.

The literature search followed the same approach as the previous meta-analysis. We used the keywords *Big Five* or *Five Factor Model* in combination with *computer-based*, *computerized*, *web-based*, or *internet-based* and *paper, mail,* or *postal* and identified 10 studies that met the inclusion criteria described above (Table S1). We extracted the same information from the primary studies as in the previous meta-analysis. Two effect sizes (3% of all effects) were identified as outliers. Again, a subset of nine studies was independently coded twice; the two coders showed high agreement with median values of ICC and κ of 1.00. The meta-analytic procedure followed the approach previously outlined. Again, effect sizes were computed in such a way that negative effect sizes indicated lower trait scores and, thus, less social desirability distortion on the computer.

Results

Sample characteristics. The meta-analysis pooled 66 effect sizes from a total of 2,951 participants with about 40% coming from the US. The 14 independent samples included about 70% female participants and had a mean age of 23.88 years (*SD* = 6.59). More than two thirds of the samples used student participants (71%), whereas the rest included adult employees. All studies were published in peer-reviewed journals. With regard to the administered instruments, about 38% of all effect sizes were based on the International Personality Item Pool (Goldberg, 1999) and about 23% on the NEO-Five Factor Inventory (Costa & McCrae, 1992), whereas the remaining samples administered a variety of different instruments.

Pooled effect. The pooled adjusted effect across all traits was $\Delta = 0.05$, $p = .15$ (Table 2). Although there was some variation between the five traits (Figure 1), the overall results did not indicate less socially desirable distortion on the computer. However, the effect sizes exhibited significant heterogeneity, $Q = 106.21$, $df = 65$, $p < .001$; about 35% of the total variance in the observed effects was due to random variance. In a meta-regression model the administration mode (coded -1 for proctored and 1 for unproctored settings) explained about 2% of the heterogeneity in the effect sizes and thus showed no moderation effect, $\gamma = -0.01$, $SE = 0.04$, $p = .80$. Thus, unproctored WBT did not affect overreporting of the Big Five traits as compared to proctored computerized assessments.

Sensitivity analyses. The robustness of the previously reported results was examined in a series of sensitivity analyses. First, eliminating two job applicant samples from the meta-analytic database replicated the null effect, $\Delta = 0.01$, $p = .71$. Moreover, after removing these samples the remaining heterogeneity in effect sizes (I^2 = .12) was negligible, *Q* = 57.26, *df* = 55, *p* = .39. In contrast, for the two applicant samples the effect was significant, but, contrary to our expectations, indicated slightly more overreporting on the computer, Δ = 0.18, *p* = .03. Thus, the goal of the assessment (e.g., for research purposes or job selection) might moderate any potential mode differences to some degree. Second, the six between-sample covariates explained about 72% of the remaining random variance in non-applicant samples (Table 3). However, the only relevant moderator was the respondents' origin: US samples exhibited no mode effect, $\Delta_{\text{predicted}}$ = 0.09, whereas non-US samples indicated greater overreporting of Big Five traits on the computer, $\Delta_{\text{predicted}}$ = 0.32. The magnitude of these effects is rather negligible (cf. Ferguson, 2009). Again, unproctored WBT did not reduce social desirability compared to proctored assessments in this model, γ = 0.05, *SE* = 0.05, *p* = .32 (Table 3).

Publication bias. The funnel plot (middle plot in Figure 2) showed a largely symmetrical distribution of the effect sizes around the population effect. Neither the rank correlation test, τ = -.02, *p* = .83, nor the regression test for funnel plot asymmetry, γ = -0.59, *SE* = 0.52, *p* = .26, indicated any potential publication bias.

Meta-Analysis III: Psychopathology

Method

The third meta-analysis focused on computer- and paper-and-pencil-administered measures of psychopathological symptoms. We assumed that individuals generally strive to appear well-adjusted and healthy; thus, lower levels of reported psychopathological symptoms would be indicative of socially desirable responses (Baer & Miller, 2002; McGrath, Mitchell, Kim, & Hough, 2010). Prevalent models of psychological disorders (cf. Krueger & Markon, 2006) classify mental disorders into two broad clusters: the cluster of "internalizing" disorders including distress disorders (e.g., depression and anxiety) and fear disorders (e.g., phobias), and the cluster of "externalizing" disorders (e.g., substance-use disorders). In line with this classification, the meta-analysis focused on four symptom groups:

(a) depressive symptoms, (b) generalized anxiety symptoms, (c) phobic symptoms, and (d) symptoms related to alcohol and drug use.

Following the same meta-analytical procedure as in the previous studies, we pooled effects from 28 studies identified from a literature search using the keywords *depression*, *anxiety, phobia, alcohol-dependency,* or *drug use* in combination with *computer-based*, *computerized*, *web-based*, or *internet-based* and *paper, mail,* or *postal* (Table S1). As in the previous meta-analyses, negative effects indicated higher levels of psychopathology and thus less social desirability distortion on the computer. One effect size (1% of all effects) was classified as an outlier. Two independent ratings of a subsample of 11 studies showed high agreement with a median ICC and Cohen's κ of 1.00 [0.93, 1.00].

Results

Sample characteristics. The meta-analysis was comprised of 96 effect sizes from a total of 16,034 participants (64% female) with a mean age of 32.51 years (*SD* = 11.25). Twenty-six percent of the studies were conducted on patients with psychological disorders seeking treatment, whereas about 28% reported on student samples. About 36% of the studies were conducted in the US. All but one study were published in peer-reviewed journals.

Pooled effect. The pooled adjusted effect of computerized assessments on socially desirable responding across all four symptom groups was Δ = 0.00, *p* = .87 (Table 2) and thus identified no mode effect. Although there was some variation in the pooled effects across the symptom groups (Figure 1), there was no evidence of less socially desirable distortions in the computerized tests. Moreover, proctored and unproctored assessments did not yield different results, γ = 0.03, *SE* = 0.02, *p* = .15. Thus, WBT did not increase reports of psychopathological symptoms. Overall, there was little heterogeneity in the effects, *Q* = 84.24, *df* = 95, *p* = .78 (Table 2); only about 17% of the total variance in the observed effects was due to random variance.

Sensitivity analyses. First, we examined whether the assessment mode showed stronger effects for samples including patients with psychological disorders than for community samples. However, a meta-regression analysis did not support this assumption, $\gamma = 0.03$, $SE = 0.02$, $p = .26$. Thus, the sample type did not moderate potential mode effects. Second, we studied the influence of the same between-sample covariates as in the previous meta-analyses (Table 3). These analyses revealed significantly stronger mode differences for between-subject designs, $\Delta_{\text{predicted}} = -0.08$, than for within-subject designs, $\Delta_{\text{predicted}} = 0.02$. More importantly, after controlling for the other moderators, the assessment setting also had a significant impact on self-reports of psychopathological symptoms. In contrast to our expectations, unproctored assessments, $\Delta_{\text{predicted}} = 0.01$, resulted in less mode differences than proctored assessments, $\Delta_{\text{predicted}} = -0.07$. Thus, these results offer no support for the candor hypothesis.

Publication bias. Because only one unpublished study was available, we refrained from comparing effect sizes from published and unpublished sources. The funnel plot for the entire meta-analytic database (right plot in Figure 2) showed a fairly symmetrical distribution of the observed effect sizes and thus no evidence for a publication bias. Moreover, neither the rank correlation test, $\kappa = -.07$, $p = .32$, nor the regression test for funnel plot asymmetry, $B = -0.43$, $SE = 0.25$, $p = .09$, indicated a publication bias. Thus, the presented results do not seem to be distorted by a publication bias.

Discussion

Research exploring mode differences in survey designs is extensive and continues to grow (cf. Couper, 2011; Gnambs & Kaspar, 2014). One dominating topic in this field pertains to the question of whether certain assessment modes are associated with specific response styles. According to the “candor” hypothesis (Buchanan, 2000, 2001) WBT results in responses showing less socially desirable distortion than comparable paper-and-pencil surveys. This hypothesis has sparked a number of survey experiments that, so far, have provided rather heterogeneous results. Therefore, this study aimed to consolidate this area of research and provide a comprehensive summary of available empirical findings. This led to the conclusion that the administration mode did affect neither self-reported social desirability (Meta-Analysis I), nor overreporting of favorable personality characteristics (Meta-Analysis II), nor underreporting of mental health symptoms (Meta-Analysis III). Overall, all pooled effects were rather low (between -.11 and .09, see Figure 1) and were far from being of practical relevance (Ferguson, 2009). Thus, these results do not support the "candor" hypothesis (Buchanan, 2000, 2001). Apparently, computerized testing alone, even in the form of unproctored WBT, is not sufficient for people to reduce impression management tactics. Given the current global debate on data security on the Internet and a gradually growing awareness of one's limited privacy when being online, it seems unlikely that response distortions will evolve over the coming years in such a way as to support the predictions of the "candor" hypothesis.

The Future of the "Candor" Hypothesis

Although it might be tempting to completely dismiss the "candor" hypothesis in light of the presented meta-analytic findings, this conclusion might be premature. A recent meta-analysis on self-disclosure of sensitive behaviors (Gnambs & Kaspar, 2014) estimated that respondents were about 1.5 times more likely to admit to various controversial behaviors such as drug use and various sexual activities when interviewed on a computer as compared to paper-and-pencil. Moreover, these survey mode distortions were most pronounced for the most sensitive behaviors. Although the study did not explicitly focus on WBT, its results show that survey mode differences can affect self-reports in some situations. Thus, one might speculate that a reason for the present null findings relates to the examined topics: some self-reports may be more likely to elicit social desirability distortions than others. Highly sensitive topics that are more susceptible to social judgments (e.g., on sexual well-being or psychopathic tendencies) may be more strongly affected by survey mode differences than measures of impression management, personality, or depression.

Also, survey modes might not comparably affect respondents in all situations. For example, in student and general community samples that complete psychological tests for research purposes without having to fear individual consequences, survey modes do not seem to distort self-reports (e.g., Chuah et al., 2006; Weigold et al., 2013), a result confirmed by the presented meta-analyses. Although psychiatric patients showed similar results in our third meta-analysis, one may speculate that some people with specific psychological disorders, such as social anxieties, could benefit from computerized assessments: Computers are frequently perceived as neutral and anonymous communicators (Buchanan, 2000; Joinson, 1999). Moreover, when dealing with computers, people are often completely immersed in the task at hand (cf. the concept of transportation; Gnambs, Appel, Schreiner, Richter, & Isberner, 2014). As a consequence, computerized assessments might put less pressure on socially anxious patients to respond in line with socially approved norms.

Finally, in situations where people are motivated to misrepresent themselves, such as during job selection processes, several studies showed that applicants even tend to slightly overreport personality traits (e.g., Ployhart et al., 2003; Salgado & Moscoso, 2003)—an effect that was replicated in our second meta-analysis (albeit based on only two samples). Although the reasons for these context-specific mean-level discrepancies have not yet been fully explored, it leaves room for some intriguing speculations: Deliberate impression management is a common strategy in many online interactions; for example, people typically strive to show their most favorable selves on interactive web-platforms such as Facebook (Zhao, Grasmuck, & Martin, 2008). Hence, associations might evolve that implicitly connect computers and web-based conduct with the use of impression management tactics. Particularly in situations that result in overly positive self-presentation anyway (e.g., in selection contexts), computerized assessments may contribute to the overreporting of favorable personality characteristics. Therefore, a fruitful avenue for future research would be the identification of specific assessment goals that might moderate survey mode differences.

Implications for Psychological Assessment

The consequences of the presented meta-analyses for psychological practice are twofold. On the positive side, the results reinforce the trustworthiness of WBT (including mixed-mode designs) by establishing mean-level equivalence across survey modes. In line with related studies that documented factorial invariance (e.g., Chuah et al. 2006; Swahney & Cigularov, 2014), comparable reliabilities (e.g., Bjorner et al., 2014), and validities (Beaty et al., 2011), this study highlighted the fact that even scalar invariance can most likely be achieved for many psychological self-report scales. However, these results should not imply that mean-level equivalence can be taken for granted in all situations. For example, the diffusion of agent-based human-computer interfaces will most likely make traditional written questionnaires increasingly less prevalent in the future, when more realistic interviews using virtual agents will dominate (cf. Baur, Damian, Gebhard, Porayska-Pomsta, & André, 2013; Friederichs, Bolman, Oenema, Guyaux, & Lechner, 2014). Preliminary evidence indicates that people report lower impression management and display more sadness when being interviewed by a fully automated virtual human as compared to computerized assessments involving interactions with real humans (Lucas, Grath, King, & Morency, 2014).

On the negative side, the present results revealed *no advantages* of WBT regarding socially desirable distortions. Rather, computerized and paper-and-pencil questionnaires seem to be comparably affected by social desirability biases in proctored and unproctored test settings. Thus, any hopes that a switch from paper to computer would automatically improve psychological measurements (Buchanan, 2000, 2001) need to be abandoned. However, computerized testing (including unproctored WBT) holds a variety of additional advantages, such as the use of simulation-based assessment scenarios (Schönbrodt & Asendorpf, 2011) or adaptive item presentations (Gnambs & Batinic, 2011; Simms et al. 2011).

Limitations

Some limitations of this work must be noted. First, it might be speculated that differences in the psychometric properties of the administered instruments might have distorted any potential mode effects. The meta-analyses relied on reported sample statistics to infer social desirability effects. Comparisons of these values require measurement equivalence across presentation modes; that is, paper-and-pencil and computerized testing need to measure the same construct in a comparable way to draw valid inferences from the observed mean statistics. Although the examination of measurement invariance is beyond the scope of this study, previous mode comparisons confirmed measurement invariance of self-report measures across media (Bjorner et al., 2014; Chuah et al., 2006; Weigold et al., 2013).

Second, survey modes might have also affected other response styles such as acquiescence, extreme or midpoint responding that have not been acknowledged in this study. Indeed, preliminary findings suggest that there are small differences in these response styles between survey modes (Weijters, Schillwaert, & Geuens, 2008). Respondents seem to engage in more acquiescence reporting in some variants of WBT as compared to postal surveys; that is, people tend to agree with items in the former mode more readily than in the latter. Thus, if scales are administered that do not include reverse-scored items, acquiescence responding would result in higher means in WBT, which could be misinterpreted as resulting from less socially desirable responding. Thus, future mode experiments should use a set of items including positive and negative wording.

Third, only few studies examined mode effects in applied settings; most available studies focused on student samples. Therefore, little is known about the status of the "candor" hypothesis in situations where the test outcome matters for the respondents because it determines, for example, job selection decisions or psychiatric diagnoses.

Finally, the meta-analyses focused on the survey medium and one aspect of the administration setting (proctored versus unproctored). However, testing situations can vary along a variety of dimensions that might differently influence social desirability, such as the presence of third persons during the interview or the use of different technologies (e.g., home computers or smartphones). Therefore, it appears necessary to further explore the psychological mechanisms that might trigger effects of the administration mode. For example, computerized testing may lead to a stronger feeling of anonymity by reducing socioemotional nonverbal cues and personal characteristics (e.g., one's gender or cultural background). This, in turn, can elicit deindividuation tendencies that lower the threshold for norm violations and less socially desirable behavior. Apparently, this is the prevailing view stated in the literature (cf. Buchanan, 2000; Tourangeau & Yan, 2007). Alternatively, it is also conceivable that the critical point is the extent to which participants believe in their identifiability (cf. Joinson & Paine, 2006). In fact, the test setting can be very anonymous (independent of administration mode) while the test subject remains nonetheless identifiable—for example, by means of their IP address, written informed consent, or a participant code. Thus, anonymity and identifiability are two distinct concepts (for a theoretical framework see the SIDE model in computer-mediated communication, Reicher, Spears, & Postmes, 1995). Therefore, future mode experiments should scrutinize additional factors that might affect WBT beyond mere comparisons with traditional media.

Conclusions

In sum, the research presented here shows no support for the "candor" hypothesis (Buchanan, 2000, 2001). Three meta-analyses concordantly failed to identify less social desirable responding in web-based as compared to paper-and-pencil surveys. Overall, social desirability in self-report scales does not seem to be affected by the adopted survey mode. These results provide further confidence in the use of web-based assessments and mixed-mode designs in survey research.

Footnotes

[1] It should be noted that we had no a priori hypotheses regarding potential effects of these variables. We coded them because we expected that relevant information would be reported in most primary studies and, thus, would allow for detailed sensitivity analyses of the pooled effect across a variety of conditions.

[2] The sample type was not included as a moderating variable because all but one study reported on student samples.

References

Aluja, A., Rossier, J., & Zuckerman, M. (2007). Equivalence of paper and pencil vs internet forms of the ZKPQ-50-CC in Spanish and French samples. *Personality and Individual Differences*, *43*, 2022-2032. doi:10.1016/j.paid.2007.06.007

Arbeitskreis Deutscher Marktforschungsinstitute (ADM). (2015). *Marktforschung in Zahlen 2/2015* [Market research in numbers]. https://www.adm-ev.de/zahlen/

#Austin, D. W., Carlbring, P., Richards, J. C., & Andersson, G. (2006). Internet administration of three commonly used questionnaires in panic research: Equivalence to paper administration in Australian and Swedish samples of people with panic disorder. *International Journal of Testing, 6*, 25-39. doi:10.1207/s15327574ijt0601_2

Baer, R. A., & Miller, J. (2002). Underreporting of psychopathology on theMMPI-2: A meta-analytic review. *Psychological Assessment, 14*, 16-26. doi:10.1037/1040-3590.14.1.16

Bartram, D. (2000). Internet recruitment and selection: Kissing frogs to find princes. *International Journal of Selection and Assessment, 8*, 261-274. doi:10.1111/1468-2389.00155

Beaty, J. C., Nye, C. D., Borneman, M. J., Kantrowitz, T. M., Drasgow, F., & Grauer, E. (2011). Proctored versus unproctored Internet tests: Are unproctored noncognitive tests as predictive as job performance? *International Journal of Selection and Assessment, 19,* 1-10. doi:10.1111/j.1468-2389.2011.00529.x

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088-1101. doi:10.2307/2533446

#Beebe, T. J., Harrison, P. A., Park, E., McRae, J. A. Jr., & Evans, J. (2006). The effects of data collection mode and disclosure on adolescent reporting of health behavior. *Social Science Computer Review, 24*, 476-488. doi:10.1177/0894439306288690

+Bjornsdottir, G., Almarsdottir, A. B., Hansdottir, I., Thorsdottir, F., Heimisdottir, M., Stefansson, H., …, Brennan, P. F. (2014). From paper to web: Mode equivalence of the ARHQ and NEO-FFI. *Computers in Human Behavior, 41*, 384-392. doi:10.1016/j.chb.2014.10.033

Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E. (2014). Difference in method of administration did not significantly impact item response: an IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. *Quality of Research, 23*, 217-227. doi:10.1007/s11136-013-0451-4

Boeker, M., Vach, W., & Motschall, E. (2013). Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC Medical Research Methodology*, *13*:131. doi:10.1186/1471-2288-13-131

*Booth-Kewley, S., Larson, G. E., & Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior, 23*, 463-477. doi:10.1016/j.chb.2004.10.020

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*, 281-291. doi:10.1093/pubmed/fdi031

Brannick, M. T., Yang, L.-Q., & Cafri, G. (2011). Comparison of weights for meta-analysis of *r* and *d* under realistic conditions. *Organizational Research Methods*, *14*, 587-607. doi:10.1177/1094428110368725

#Broering, J. M., Paciorek, A., Carroll, P. R., Wilson, L. S., Litwin, M. S., & Miaskowski, C. (2014). Measurement equivalence using a mixed-mode approach to administer health-related quality of life instruments. *Quality of Life Research, 23*, 495-508. doi:10.1007/s11136-013-0493-7

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

Buchanan, T. (2001). Online personality assessment. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science* (pp. 57-74). Lengerich, Germany: Pabst Science Publishers.

#Carlbring, P., Brunt, S., Bohman, S., Austin, D. W., Richards, J. C., Öst, L. G., & Andersson, G. (2007). Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior*, *23*, 1421-1434. doi:10.1016/j.chb.2005.05.002

#Caro, J. J., Caro, I., Caro, J., Wouters, F., & Juniper, E. F. (2001). Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? *Quality of Life Research, 10*, 683-691. doi:10.1023/A:1013811109820

#Chen, T.-H., Li, L., Sigle, J. M., Du, Y.-P., Wang, H.-M., & Lei, J. (2007). Crossover randomized controlled trial of the electronic version of the Chinese SF-36. *Journal of Zhejiang University SCIENCE B, 8*, 604-608. doi:10.1631/jzus.2007.B0604

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*, 211-229. doi:10.1037/a0032968

Cheung, M. W.-L. (2015). metaSEM: An *R* package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*:1521. doi:10.3389/fpsyg.2014.01521

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine, 19*, 3127-3131. doi:10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M

+Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality, 40*, 359-376. doi:10.1016/j.jrp.2005.01.006

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101-129. doi:10.2307/3001666

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

Couper, M. P. (2011). The future of modes of data collection. *Public Opinion Quarterly, 75*, 889-908. doi:10.1093/poq/nfr046

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354. doi:10.1037/h0047358.

De Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as part of a mixed mode design. In M. Das, P. Ester, & L. Kaczmirek (Eds), *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 45-76). New York, NY: Taylor & Francis.

Dwight, S. A., & Feigelson, M. E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement, 60*, 340-360. doi:10.1177/00131640021970583

#Eaton, D. K., Brener, N. D., Kann, L., Denniston, M. M., McManus, T., Kyle, T. M., …, & Ross, J. G. (2010). Comparison of paper-and-pencil versus web administration of the youth risk behaviour survey (YRBS): Risk behavior prevalence estimates. *Evaluation Review, 34*, 137-153. doi:10.1177/0193841X10362491

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634. doi:10.1136/bmj.315.7109.629

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532-538. doi:10.1037/a0015808

Fogarty, T. J., Jonas, G. A., & Parker, L. M. (2013). The medium is the message: Comparing paper-based and web-based course evaluation modalities. *Journal of Accounting Education, 31*, 177-193. doi:10.1016/j.jaccedu.2013.03.002

*Fox, S., & Schwartz, D. (2002). Social desirability and controllability in computerized and paper-and-pencil personality questionnaires. *Computers in Human Behavior, 18*, 389-410. doi:10.1016/S0747-5632(01)00057-7

Friederichs, S., Bolman, C., Oenema, A., Guyaux, J., & Lechner, L. (2014). Motivational interviewing in a web-based physical activity intervention with an avatar: Randomized controlled trial. *Journal of Medical Internet Research*, *16*:e48. doi:10.2196/jmir.2974

Galletly, C. L., & Pinkerton, S. D. (2006). Conflicting messages: How criminal HIV disclosure laws undermine public health efforts to control the spread of HIV. *AIDS and Behavior, 10*, 451-461. doi:10.1007/s10461-006-9117-3

+Gaudron, J.-P. (2000). The effects of computer anxiety on self-description with a computerized personality inventory. *European Review of Applied Psychology, 50*, 431-436.

*Gerich, J. (2008). Real or virtual? Response behavior in video-enhanced self-administered computer interviews. *Field Methods, 20*, 356-376. doi:10.1177/1525822X08320057

Gnambs, T. (2013). The elusive general factor of personality: The acquaintance effect. *European Journal of Personality, 27*, 507-520. doi:10.1002/per.1933

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality, 52*, 20-28. doi:10.1016/j.jrp.2014.06.003

Gnambs, T., & Batinic, B. (2011). Polytomous adaptive classification testing effects of item pool size, test termination criterion, and number of cutscores. *Educational and Psychological Measurement*, *71*, 1006-1022. doi:10.1177/0013164410393956

Gnambs, T., Appel, M., Schreiner, C., Richter, T., & Isberner, M.-B. (2014). Experiencing narrative worlds: A latent state-trait analysis. *Personality and Individual Differences, 69*, 187-192. doi:10.1016/j.paid.2014.05.034

Gnambs, T., & Kaspar, K. (2014). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-014-0533-4

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. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg: Tilburg University Press.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877-902. doi:10.1146/annurev-psych-010814-015321

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

#van Griensven, F., Naorat, S., Kilmarx, P. H., Jeeyapant, S., Manopaiboon, C., Chaikummao, S., …, & Tappero, J. W. (2006). Palmtop-assisted self-interviewing for the collection of sensitive behavioral data: Randomized trial with drug use urine testing. *American Journal of Epidemiology, 163*, 271-278. doi:10.1093/aje/kwj038

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. Hoboken, NJ: Wiley.

#Gudbergsen, H., Bartels, E. M., Krusager, P., Waeæhrens, E. E., Christensen, R., Danneskiold-Samsøe, B., & Bliddal, H. (2011). Test-retest of computerized health status questionnaires frequently used in the monitoring of knee osteoarthritis: A randomized crossover trial. *BMC Musculoskeletal Disorders, 12*, 190-198. doi:10.1186/1471-2474-12-190

*Hancock, D. R., & Flowers, C. P. (2001). Comparing social desirability responding on World Wide Web and paper-administered surveys. *Educational Technology, Research and Development, 49*, 5-13. doi:10.1007/BF02504503

Hathaway, S. R., & McKinley, J. C. (1989). *Minnesota Multiphasic Personality Inventory*. New York, NY: University of Minnesota Press.

#Hayes, J., & Grieve, R. (2013). Faked depression: Comparing malingering via the Internet, pen-and-paper, and telephone administration modes. *Telemedicine and e-Health, 19*, 714-716. doi:10.1089/tmj.2012.0278

*#Hays, S., & McCallum, R. S. (2005). A comparison of the pencil-and-paper and computer-administrated Minnesota Multiphasic Personality Inventory-Adolescent. *Psychology in the Schools, 42*, 605-613. doi:10.1002/pits.20106

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128. doi:10.3102/10769986006002107

#Herrero, J., & Menese, J. (2006). Short web-based versions of the perceived stress (PSS) and Center for Epidemiological Studies-Depression (CESD) 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

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557-560. doi:10.1136/bmj.327.7414.557

#Holländare, F., Askerlund, A.-M., Nieminen, A., & Engström, I. (2008). Can the BDI-II and MADRS-S be transferred to online use without affecting their psychometric properties? *E-Journal of Applied Psychology*, *4*, 63-65. doi:10.7790/ejap.v4i2.122

#Holländare, F., Andersson, G., & Engström, I. (2010). A comparison of psychometric properties between Internet and paper versions of two depression instruments (BDI-II and MADRS-S) administered to clinic patients. *Journal of Medical Internet Research, 12* (e49). doi:10.2196/jmir.1392

John, L. K., Acquisti, A., & Loewenstein, G. (2011). Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of Consumer Research, 37*, 858-873. doi:10.1086/656423

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

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, *31*, 177-192. doi:10.1002/ejsp.36

Joinson, A. N., & Paine, C. (2006). Self-disclosure, privacy and the Internet. In A. Joinson, K. McKenna, T. Postmes, & U.-D. Reips (Eds.), *The Oxford Handbook of Internet Psychology* (pp. 237-252). Oxford, England: University Press.

Joinson, A. N., Woodley, A., & Reips, U.-D. (2007). Personalization, authentication and self-disclosure in self-administered Internet surveys. *Computers in Human Behavior*, *23*, 275-285. doi:10.1016/j.chb.2004.10.012

#Källmén, H., Sinadinovic, K., Berman, A. H., & Wennberg, P. (2011). Risky drinking of alcohol in Sweden: A randomized population survey comparing web- and paper-based self-reports. *Nordic Studies on Alcohol and Drugs, 28*, 123-130. doi:10.2478/v10199-011-0013-4

Kays, K., Gathercoal, K., & Burhow, W. (2012). Does survey format influence self-disclosure on sensitive question items? *Computers in Human Behavior, 28*, 251-256. doi:10.1016/j.chb.2011.09.007

#Kongsved, S. M., Basnov, M., Holm-Christensen, K., & Hjollund, N. H. (2007). Response rate and completeness of questionnaires: A randomized study of Internet versus paper-

and-pencil versions. *Journal of Medical Internet Research, 9*, e25. doi:10.2196/jmir.9.3.e25

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2*, 111-133. doi:10.1146/annurev.clinpsy.2.022305.095213

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*, 201-228. doi:10.1111/j.1744-6570.2009.01136.x

Lucas, G. M., Gratch, J., King, A., & Morency, L. P. (2014). It's only a computer: virtual humans increase willingness to disclose. *Computers in Human Behavior, 37*, 94-100. doi:10.1016/j.chb.2014.04.043

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*,450-470. doi:10.1037/a0019216

*Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*, 322-345. doi:10.1177/10944281062893932007

#Miller, E. T., Neal, D. J., Roberts, L. J., Baer, J. S., Cressler, S. O., Metrik, J., & Marlatt, G. A. (2002). Test-retest reliability of alcohol measures: Is there a difference between Internet-based assessment and traditional methods? *Psychology of Addictive Behaviors, 16*, 56-63. doi:10.1037/0893-164X.16.1.56

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105-125. doi:10.1037/1082-989X.7.1.105

+#Morrison-Beedy, D., Carey, M. P., & Tu, X. (2006). Accuracy of audio computer-assisted self-interviewing (ACASI) and self-administered questionnaires for the assessment o of sexual behavior. *AIDS Behavior, 10*, 541-552. doi:10.1007/s10461-006-9081-y

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61-87). San Diego, CA: Academic Press.

+Parma, M., Gali, Z., & Željko, J. (2007). On-line personality assessment: Are electronic versions equivalent to the traditional one? In V. C. Adorić (Ed.), *15th Psychology Days in Zadar* (pp. 259-272). Zadar, Croatia: University of Zadar.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609. doi:10.1037/0022-3514.46.3.598

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Paunonen, S. V., & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology, 103*, 158-175. doi:10.1037/a0028165

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

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org

#Raat, H., Mangunkusumo, R. T., Landgraf, J. M., Kloeck, G., & Brug, J. (2007). Feasibility, reliability, and validity of adolescent health status measurement by the Child Health Questionnaire Child Form (CHQ-CF): Internet administration compared with the standard paper version. *Quality of Life Research, 16*, 675-685. doi:10.1007/s11136-006-9157-1

$^+$Rammstedt, B., Holzinger, B., & Rammsayer, T. (2004). Zur Äquivalenz der Papier-Bleistift- und einer computergestützten Version des NEO-Fünf-Faktoren-Inventars (NEO-FFI) [On the equivalence of a paper-and-pencil and computerized version of the NEO-FFI]. *Diagnostica, 50*, 88-97. doi:10.1026/0012-1924.50.2.88

Reicher, S., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology, 6*, 161-198. doi:10.1080/14792779443000049

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 interview. *Journal of Applied Psychology, 84*, 754-775. doi:10.1037/0021-9010.84.5.754

$^\#$Richter, J. G., Becker, A., Koch, T., Nixdorf, M., Willers, R., Monser, R., …, Schneider, M. (2008). Self-assessments of patients via Tablet PC in routine patient care: comparison with standardised paper questionnaires. *Annals of the Rheumatic Diseases, 67*, 1739-1741. doi:10.1136/ard.2008.090209

Risko, E. F., Quilty, L. C., & Oakman, J. M. (2006). Socially desirable responding on the web: Investigating the candor hypothesis. *Journal of Personality Assessment, 87*, 269-276. doi:10.1207/s15327752jpa8703_08

*Rossiter, J. C. (2009). *A comparison of social desirability bias among four widely used methods of data collection as measured by the impression management subscales of the Balance Inventory of Desirable Responding* (Doctoral dissertation, Kent State University). Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=kent1240263500

#Ryan, J. M., Corry, J. R., Attewell, R., & Smithson, M. J. (2002). A comparison of an electronic version of the SF-36 General Health Questionnaire to the standard paper version. *Quality of Life Research, 11*, 19-26. doi:10.1023/A:1014415709997

+Salgado, J., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assesses' perceptions and reactions. *International Journal of Selection and Assessment, 11*, 194-205. doi:10.1111/1468-2389.00243

#SAMHSA (2001). *Development of computer-assisted interviewing procedures for the National Household Survey on Drug Abuse*. Substance Abuse and Mental Health Services Administration (SAMHSA), Department of Health and Human Services, Rockville, MD.

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality &Quantity, 31*, 385-399. doi:10.1023/A:1004298118485

Saville, P., Holdsworth, R., Nyfiled, G., Cramp, L., & Mabey, W. (1996). *Occupational Personality Questionnaire: Manual and user's guide*. Boston, MA: SHL.

#Schmitz, N., Hartkamp, N., Brinschwitz, C., Michalek, S., & Tress, W. (2000). Comparison of the standard and the computerized versions of the Symptom Check List (SCL-90-R): a randomized trial. *Acta Psychiatrica Scandinavica, 102*, 147-152. doi:10.1034/j.1600-0447.2000.102002147.x

Schönbrodt, F. D., & Asendorpf, J. B. (2011). Virtual social environments as a tool for psychological assessment: Dynamics of interaction with a virtual spouse. *Psychological Assessment*, *23*, 7-17. doi:10.1037/a0021049.

#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

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT–PD project. *Journal of Personality Assessment*, *93*, 380-389. doi:10.1080/00223891.2011.577475

Skitka, E. G., Sargis, L. J., & McKeever, W. (2013). The Internet as Psychological Laboratory revisited: Best practices, challenges, and solutions. In Y. Amichai-Hamburger (Ed.), *The social net: Understanding our online behavior* (pp. 253-270). Oxford, England: University Press.

#Surís, A., Borman, P. D., Lind, L., & Kashner, T. M. (2007). Aggression, impulsivity, and health functioning in a veteran population: equivalency and test–retest reliability of computerized and paper-and-pencil administrations. *Computers in Human Behavior, 23*, 97-110. doi:10.1016/j.chb.2004.03.038

+Swahney, G., & Cigularov, K. P. (2014). Measurement equivalence and latent mean differences of personality scores across different media and proctoring administration conditions. *Computers in Human Behavior, 36*, 412-421. doi:10.1016/j.chb.2014.04.010

#Swartz, R. J., de Moor, C., Cook, K. F., Fouladi, R. T., Basen-Enquist, K., Eng, C., & Taylor, C. L. (2007). Mode effects in the center for epidemiologic studies depression (CES-D) scale: personal digital assistant vs. paper and pencil administration. *Quality of Life Research, 16*, 803-813. doi:10.1007/s11136-006-9158-0

*Taddicken, M. (2009). Die Bedeutung von Methodeneffekten der Online-Befragung. In N. Jackob, H. Schoen, & T. Zerback (Eds.), *Sozialforschung im Internet* [Social research on the Internet] (pp. 91-108). Wiesbaden, Germany: Verlag für Sozialwissenschaften.

Tippins, N. T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 551-582. doi:10.1146/annurev-orgpsych-031413-091317

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859-883. doi:10.1037/0033-2909.133.5.859

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9*, 151-176. doi:10.1146/annurev-clinpsy-050212-185510

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

*Uriell, Z. A., & Dudley, C. M. (2009). Sensitive topics: Are there modal differences? *Computers in Human Behavior, 25*, 76-87. doi:10.1016/j.chb.2008.06.007

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science, 5*, 243-262. doi:10.1177/1745691610369465

#Vallejo, M. A., Mañanes, G., Comeche, M. I., & 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

+Vecchione, M., Alessandri, G., & Barbaranelli, C. (2012). Paper-and-pencil and web-based testing: The measurement invariance of the Big Five personality tests in applied settings. *Assessment, 19*, 243-246. doi:10.1177/1073191111419091

Viechtbauer, W., & Cheung, W. (2010). Outlier and influencer diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 110-125. doi:10.1002/jrsm.11

Vinciarelli, A., & Mohammadi, G. (2014). A survey of personality computing. *IEEE Transactions on Affective Computing*, *5*, 273-291. doi:10.1109/TAFFC.2014.2330816

de Vries, R. E., Zettler, I., & Hilbig, B. E. (2014). Rethinking trait conceptions of social desirability scales: Impression management as an expression of honesty-humility. *Assessment,21*, 286-299. doi:10.1177/1073191113504619

*+Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods, 18*, 53-70. doi:10.1037/a0031607

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*, 409-422. doi:10.1007/s11747-007-0077-6

*Whelan, T. J. (2008). *Effects of survey modality and access controls on perceived anonymity and socially desirable responding* (Master's thesis, North Carolina State University). Retrieved from http://www.lib.ncsu.edu/resolver/1840.16/2828

*Wilkerson, J. M., Nagao, D. H., & Martin, C. L. (2002). Socially desirable responding in computerized questionnaires: When questionnaire purpose matters more than the mode. *Journal of Applied Social Psychology, 32*, 544-559. doi:10.1111/j.1559-1816.2002.tb00229.x

#Yu, S.-C., & Yu, M. N. (2007). Comparison of Internet-based and paper-based questionnaires in Taiwan using multisample invariance approach. *CyberPsychology & Behavior, 10*, 501-507. doi:10.1089/cpb.2007.9998

Zhao, S., Grasmuck, S., & Martin, J. (2008). Identity construction on Facebook: Digital empowerment in anchored relationships. *Computer in Human Behavior, 24*, 1816-1836. doi:10.1016/j.chb.2008.02.012

* Included in the meta-analysis of social desirability scales.

+ Included in the meta-analysis of Big Five scales.

# Included in the meta-analysis of psychopathology scales.

Table 1.

Descriptive Statistics for Moderators

			Mdn / %	1.	2.	3.	4.	5.	6.	*Mdn / %*	
1.	Survey year	*Meta-analysis of social desirability scales*	2007		.45	-.48	.39	-.47	.57*	2007	*Meta-analysis of Big Five scales*
2.	Country										
	1 = United States		82%	.06		.02	-.44	-.75*	.17	40%	
	-1 = other		18%							60%	
3.	Sex (percent female)		62	.38	.30		-.35	.03	-.19	70	
4.	Age (in years)		21	-.40	.07	.28		.27	.46	21	
5.	Research design										
	1 = between-subject		88%	.02	-.31	-.01	-.24		-.03	64%	
	-1 = within-subject		12%							36%	
6.	Administration setting										
	1 = proctored		53%	.25	.13	.46	.42	.02		79%	
	-1 = unproctored		47%							21%	
1.	Survey year	*Meta-analysis of psychopathology scales*	2007								
2.	Country										
	1 = United States		36%	-.19							
	-1 = other		64%								
3.	Sex (percent female)		64	-.39*	-.15						
4.	Age (in years)		31	.07	-.51*	.08					
5.	Research design										
	1 = between-subject		59%	.09	-.14	-.21	.44*				
	-1 = within-subject		41%								
6.	Administration setting										
	1 = proctored		54%	.43*	-.37*	-.19	.21	.14			
	-1 = unproctored		46%								

* $p < .05$

Table 2.

Meta-Analysis of Socially Desirable Responding in Web-Based Questionnaires

				Observed effect		Adjusted effect		Homogeneity of effects		
	k_1	k_2	N	g	SD_g	Δ	$SE_Δ$	Q	df	I^2
Social desirability	30	17	3,746	0.01	0.21	0.03	0.03	24.76	29	.04
Self-deceptive enhancement	6	5	1,699	0.00	0.14	0.05	0.07	1.55	5	.00
Impression management	22	15	3,568	0.01	0.24	0.02	0.04	22.98	21	.07
Big Five	66	14	2,951	0.02	0.20	0.05	0.03	106.21*	65	.35
Conscientiousness	15	14	2,519	0.02	0.18	0.05	0.04	11.32	14	.00
Agreeableness	13	12	2,417	0.01	0.21	0.09	0.05	25.21*	12	.48
Emotional Stability	13	12	2,417	0.08	0.18	0.08	0.06	32.64*	12	.56
Openness	12	11	1,929	-0.03	0.12	0.01	0.05	10.65	11	.16
Extraversion	13	12	2,859	0.06	0.17	0.05	0.06	24.36*	12	.47
Psychopathology	96	39	16,034	-0.01	0.18	0.00	0.02	84.24	95	.17
Depression	41	32	14,898	-0.03	0.17	0.00	0.02	31.67	40	.05
Anxiety	6	5	3,643	-0.06	0.37	0.06	0.12	13.92*	5	.58
Phobia	27	8	3,731	0.02	0.17	0.04	0.03	13.25	26	.00
Substance dependencies	12	3	171	0.03	0.09	-0.11	0.09	20.34	21	.38
Overall	184	62	21,896	0.01	0.19	0.01	0.02	217.29*	183	.24

Note. k_1 = Number of effect sizes; k_2 = Number of samples; N = Total sample size; g = Pooled unweighted standardized difference; Δ = Pooled inverse variance-weighted standardized difference; $SE_Δ$ = Standard error of Δ; Q = Test for homogeneity of effect sizes (Cochran, 1954); I^2 = Proportion of total variance in observed effects due to random variance (Higgins et al., 2003); Effect sizes are negative when there was less social desirability distortion on the computer and positive when there was more social desirability distortion on the computer.

* $p < .05$

Table 3.

Moderator Analyses of Social Desirability Effects

		Meta-Analysis I: Social desirability			Meta-Analysis II: Big Five			Meta-Analysis III: Psychopathology		
		Predicted effect	γ	*SE*	Predicted effect	γ	*SE*	Predicted effect	γ	*SE*
	Intercept (γ_0)		0.01	0.11		0.21*	0.11		-0.03	0.06
	Random level 2 variance $\tau^2_{(2)}$		0.00[a]			0.00	0.00		0.00	0.00
	Random level 3 variance $\tau^2_{(3)}$		0.00[a]			0.00[a]			0.00[a]	
1.	Publication year (γ_1)		0.01 0.01		0.01	0.01		0.00	0.01	
	Year 2004	-0.04			0.12			0.01		
	Year 2014	0.01			0.21			-0.03		
2.	Country (γ_2)		0.09*	0.05		-0.11*	0.05		0.03	0.03
	1 = United States	0.10			0.09			0.00		
	-1 = other countries	-0.08			0.32			-0.06		
3.	Sex (γ_3)		0.00	0.00		0.00	0.00		0.00	0.00
	-50 = men	0.00			0.36			-0.06		
	50 = women	0.02			0.06			0.00		
4.	Age (γ_4)		0.00	0.01		-0.01	0.01		0.00	0.00
	0 = 20 years	0.01			0.21			-0.03		
	10 = 30 years	0.02			0.06			-0.03		

Table 3. (continued)

	Meta-Analysis I: Impression Management			Meta-Analysis II: Big Five			Meta-Analysis III: Psychopathology		
	Predicted effect	γ	*SE*	Predicted effect	γ	*SE*	Predicted effect	γ	*SE*
5. Research design (γ_5)		0.02	0.05		0.01		0.05[*]	0.02	
					0.03				
1 = within-subject	0.03			0.22			0.02		
-1 = between-subject	-0.01			0.19			-0.08		
6. Administration setting (γ_6)		0.02	0.04		0.05	0.05		0.04[*]	0.02
-1 = proctored	-0.01			0.26			-0.07		
1 = unproctored	0.03					0.01			
			0.16						
Number of effect sizes (level 2)		30			56[b]			96	
Number of samples (level 3)		17			12[b]			39	

Note. Effect sizes are negative when there was less social desirability distortion on the computer and positive when there was more social desirability distortion on the computer. γ_0 = Pooled adjusted effect after correcting for moderators; γ = Fixed effects weight; *SE* = Standard error of γ; τ^2 = Random level 2 or level 3 variance of γ_0; [a] Constrained parameter. [b] Two job applicant samples were excluded from these analyses.

[*] $p < .05$

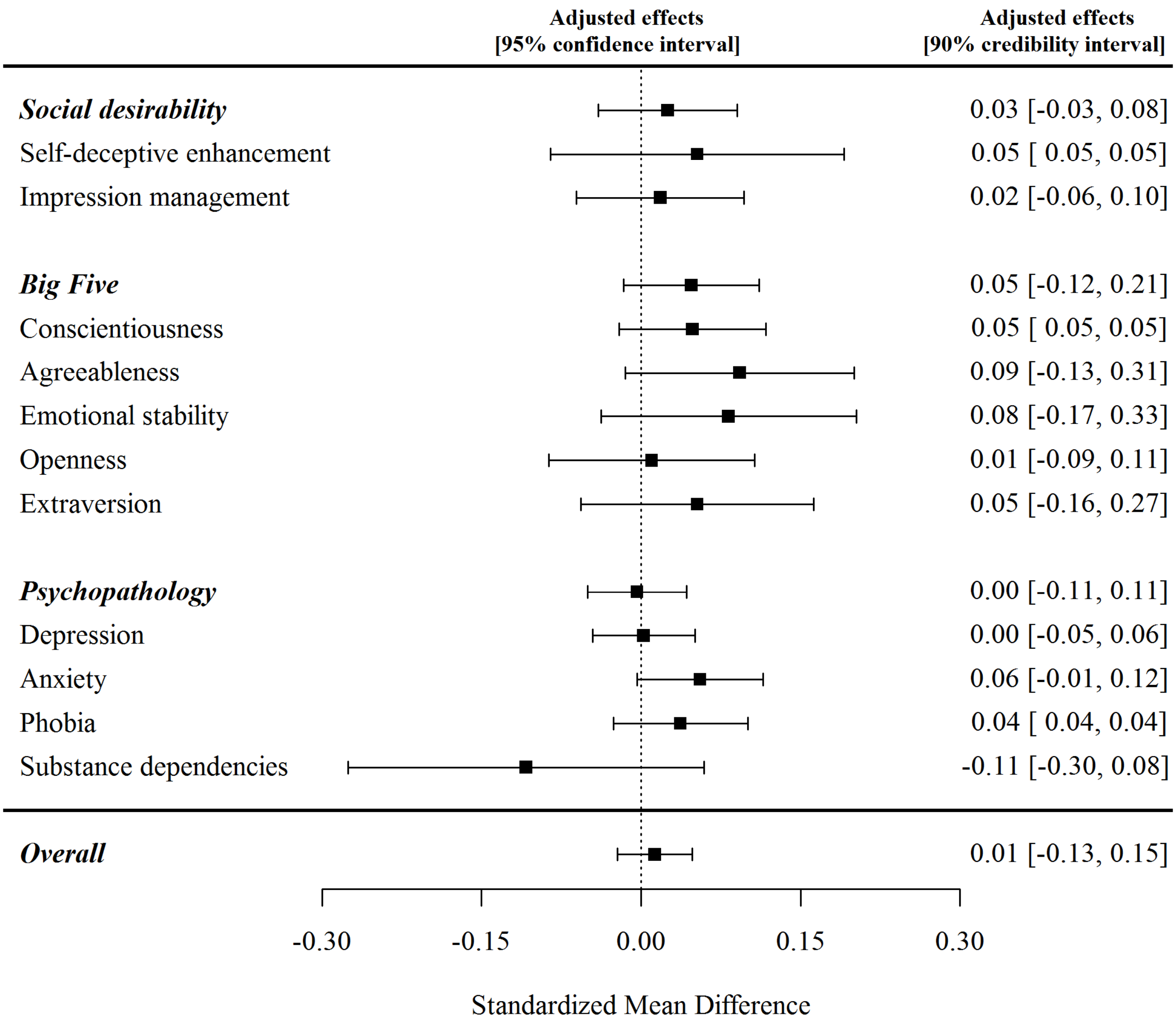

Figure 1. Forest plot for meta-analyses of standardized mean differences between computerized and paper-and-pencil assessments. Effects are negative when there was less social desirability distortion on the computer and positive when there was more social desirability distortion on the computer.

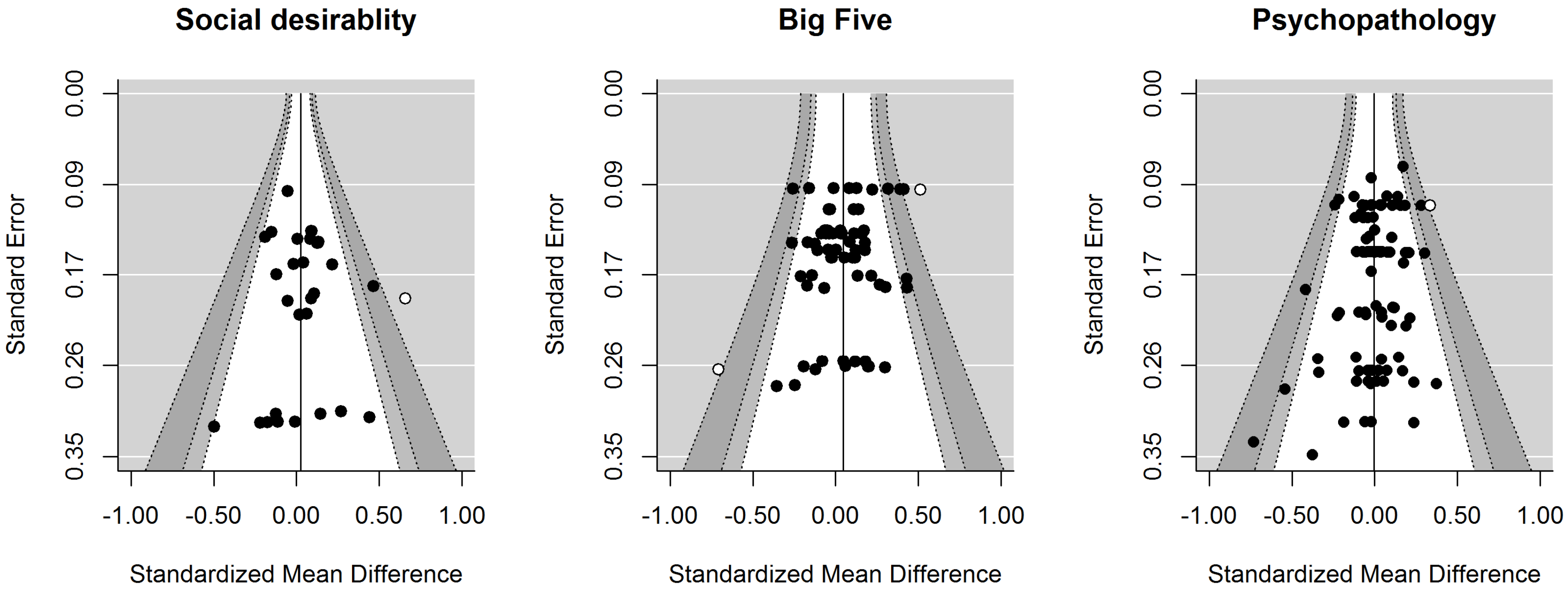

Figure 2. Contour-enhanced funnel plots for social desirability, Big Five, and psychopathology scales with 90% (white), 95% (light gray), and 99% (dark gray) confidence intervals around the pooled adjusted effect (horizontal line); white dots indicate outliers.

Socially Desirable Responding in Web-Based Questionnaires: A Meta-Analytic Review of the Candor Hypothesis

Supplement

Timo Gnambs & Kai Kaspar

Table S1.

Summary of Literature Search

	Meta-analysis I: Social desirability	Meta-analysis II: Big Five	Meta-analysis III: Psychopathology
Identified studies:			
From scientific databases	46	15	341
From Google Scholar	1,000	1,000	1,000
Excluded studies:			
Considered irrelevant after screening of title and abstract	1,033	1,002	1,310
No validated scale (criterion A)	0	0	0
Lack of randomization (criterion B)	0	2	1
Different assessment settings (criterion C)	1	1	4
Not effect size reported (criterion D)	0	0	0
Included studies:	12	10	28