Stöber, J., Dette, D. E., & Musch, J. (2002). Comparing continuous and dichotomous scoring of the Balanced Inventory of Desirable Responding. *Journal of Personality Assessment, 78*, 370-389.

Comparing Continuous and Dichotomous Scoring of The Balanced Inventory of Desirable Responding (BIDR)

Joachim Stöber* Dorothea E. Dette

Martin Luther University of Halle-Wittenberg

Jochen Musch

University of Bonn

*Author for correspondence:

Dr. Joachim Stöber
Department of Educational Psychology
Martin Luther University of Halle-Wittenberg
Franckesche Stiftungen, Haus 5
D-06099 Halle (Saale)
Germany
Phone: +49-345-5523789
Fax: +49-345-5527244
E-mail: stoeber@paedagogik.uni-halle.de

Abstract

The Balanced Inventory of Desirable Responding (BIDR) is a widely-used instrument to measure the two components of social desirability: self-deceptive enhancement (SDE) and impression management (IM). With respect to scoring of the BIDR, Paulhus (1994) has authorized two methods, namely continuous scoring (all answers on the continuous answer scale are counted) and dichotomous scoring (only extreme answers are counted). In the present article, three studies with student samples are reported, and continuous and dichotomous scoring of BIDR subscales are compared with respect to reliability, convergent validity, sensitivity to instructional variations, and correlations with personality. Across studies, the scores from continuous scoring (continuous scores) showed higher Cronbach's alphas than those from dichotomous scoring (dichotomous scores). Moreover, continuous scores showed higher convergent correlations with other measures of social desirability and more consistent effects with self-presentation instructions (fake-good versus fake-bad instructions). Finally, continuous SDE scores showed higher correlations with those traits of the five-factor model for which substantial correlations were expected (i.e., neuroticism, extraversion, and conscientiousness). Consequently, the present findings indicate that continuous scoring may be preferable to dichotomous scoring when assessing socially desirable responding with the BIDR.

Keywords: Social Desirability, Self-Deceptive Enhancement, Impression Management, Instructional Variations, Big Five Personality Traits

Comparing Continuous and Dichotomous Scoring of The Balanced Inventory of Desirable Responding (BIDR)

Since the early 1930's, the question of social desirability in self-reports has been a major concern for researchers and practitioners. Consequently, investigators have sought ways to assess social desirability and, if necessary, control for associated distortions in participants' self-reports. One major road in this endeavor was to develop scales to assess individual differences in socially desirable responding. Over the years, numerous such scales have been developed and enjoy wide application in both basic and applied research, even though the use of this practice has been repeatedly called into question (for a recent example, see Piedmont, McCrae, Riemann, & Angleitner, 2000). Nevertheless, socially desirable responding is still a prominent research topic, and continues to present a challenge to psychological measurement and personality assessment (Paulhus, in press).

Early attempts to assess socially desirable responding regarded social desirability as a one-dimensional construct. Consequently, measures to assess social desirability--such as the Edwards' Social Desirability Scale (Edwards, 1957), the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960), or the Eysenck Lie Scale (Eysenck & Eysenck, 1964) to mention just a few prominent examples--were one-dimensional in nature. However, the notion that social desirability is a unitary construct became problematic when an increasing number of studies indicated low correlations between different measures of social desirability. This problem was resolved when Paulhus (1984) inspected the correlations between various social-desirability scales and found that they formed a two-factorial space. The scales in this factor analyses included the Self-Deception Questionnaire (SDQ) and the Other Deception Questionnaire (ODQ) devised by Sackeim and Gur (1978). Because SDQ scores clearly loaded on one factor and ODQ scores clearly loaded on the

other factor, Paulhus termed the two factors "self-deceptive enhancement" and "impression management." Self-deceptive enhancement refers to an unconscious positive bias in item responses with the aim of protecting positive self-esteem. In contrast, impression management refers to the conscious dissimulation of item responses with the aim of making a favorable impression on others (Paulhus, 1986).

Sackeim and Gur's (1978) questionnaires proved unable to adequately capture these two dimensions of social desirability, however. All SDQ items are negatively keyed, so that endorsement indicates denial of negative qualities, whereas all ODQ items are positively keyed, so that endorsement indicates attribution of positive qualities. Therefore, the SDQ confounds self-deception with denial, while the ODQ confounds impression management with attribution. To deal with this shortcoming, Paulhus (1984) developed the Balanced Inventory of Desirable Responding (BIDR). Starting by rewriting some of the items from Sackeim and Gur's questionnaires and constructing new items tapping into the constructs aimed at, Paulhus arrived at a balanced inventory in which all 40 statements are affirmations, and there are equal numbers of attribution and denial items for each of the two 20-item scales measuring Self-Deceptive Enhancement (SDE) and Impression Management (IM). Continuous revision and improvement of the BIDR finally led to Version 6 of the BIDR that Paulhus began distributing in 1988 (see Paulhus, in press). Six years later, Paulhus also distributed a reference manual for BIDR Version 6 (Paulhus, 1994), giving users free access to information on reliability, convergent and discriminant validity, and norms, as well as correlations with personality and adjustment. Consequently, the BIDR Version 6 now enjoys widespread use within the scientific community, as well as great popularity in both basic and applied fields of psychology research. Though a slightly revised seventh version has recently

been published (Paulhus, 1998), Version 6 remains the most widely applied version of the BIDR.

Over the years, various studies have been conducted with the BIDR Version 6 (see Paulhus, 1994). In these studies, this version has been found to be a robust measure showing satisfactory internal consistency and test-retest reliability (Paulhus, 1991, 1994). Moreover, scores from the subscales have shown distinct validity: With respect to convergent correlations with other measures of social desirability, SDE scores have shown high correlations with scales representing the self-deceptive enhancement component of social desirability such as the Edwards' Social Desirability Scale. At the same time, IM scores have shown high convergent correlations with scales representing the impression management component of social desirability such as the Eysenck Lie Scale. (The Marlowe-Crowne Scale characteristically shows substantial correlations with both SDE and IM.) With respect to instructional variations, studies have shown that the IM scale is highly sensitive to variations of anonymity and self-presentation instructions, with participants showing substantially higher IM scores under public than private conditions, and under "fake-good" instructions than under "honest" or "fake-bad" instructions. SDE scores, in contrast, have been largely insensitive to such instructional variation. Instead, SDE scores have successfully been used to predict hindsight, overconfidence, and overclaiming (Paulhus, 1994). Moreover, SDE scores have shown substantial correlations with measures of adjustment, being associated with self-esteem and vocational identity and (inversely) with anxiety, depression, and distress. For IM scores, no substantial correlations with measures of adjustment have been found (Paulhus, 1994; Paulhus & Reid, 1991).

With respect to personality, studies have shown that SDE and IM also show different relations with personality traits, for example, when the BIDR is correlated with measures of

the "Big Five" personality traits, namely neuroticism, extraversion, openness, agreeableness, and conscientiousness (Costa & McCrae, 1992). Across studies (e.g., Meston, Heiman, Trapnell, & Paulhus, 1998; Paulhus, 1994; see also Davies, French, & Keogh, 1998) the following pattern emerged: SDE scores have shown substantial negative correlations with neuroticism and substantial positive correlations with conscientiousness. Moreover, positive correlations with extraversion have been reported, though SDE scores seem to be unrelated to agreeableness. In comparison, IM scores have shown substantial positive correlations with agreeableness and conscientiousness, while being unrelated to neuroticism and extraversion. For the trait of openness, findings on the relationship with the BIDR subscales are less consistent. Whereas some studies report substantial positive correlations of SDE with openness (e.g., Paulhus, 1994), others fail to find such relationships (e.g., Paulhus & Reid, 1991).

Despite the many studies conducted with the BIDR Version 6, one question regarding its use has been largely neglected, namely the best scoring method for the BIDR items. In the BIDR Version 6, items may be administered with two alternate answer formats, either a 5-point or a 7-point Likert-type answer scale. More importantly, Paulhus (1994) has authorized two alternate scoring methods: (a) continuous scoring and (b) dichotomous scoring. With continuous scoring, inversely keyed items are reversed, and points associated with each answer are then summed across items . With dichotomous scoring, inversely keyed items are reversed, but only extreme responses are counted. With the 5-point answer scale, one point is awarded for each "5" response on SDE items and for each "4" or "5" response on IM items. With the 7-point answer scale, one point is awarded for each "6" or "7" response on both SDE and IM items. Points are then summed across all items to form subscale scores. Paulhus (1994) gives the following recommendation as to the preferred

answer format and scoring procedure:

> Of the four scoring procedures, dichotomous scoring of 7-point scales is recommended. The stringency of the dichotomous scoring guarantees that high scores are attained only by subjects who give exaggerated responses to items that are already highly desirable. Thus, for both scales, the format seems to optimal for indexing inflated self-descriptions. (p. 7)

When reviewing the studies with the BIDR Version 6, however, we found that most authors do not follow Paulhus' (1991, 1994) recommendation. Though they use the 7-point answer scale when applying the BIDR, they do not use dichotomous scoring, but chose continuous scoring instead. There are three potential reasons for this. First, one may argue that social desirability is not an all-or-nothing process. Instead, it may be plausible to assume that the processes underlying socially desirable responding are continuously distributed variables on both the item and the composite levels (G. Becker & Cherny, 1992). Second, in counting extreme answers only, one might ignore individuals who do have a tendency for desirable responding, but at the same time, avoid extreme answers. Dichotomous scoring might thus lead to extremity bias being taken for socially desirable responding. Finally, one may argue that dichotomizing continuous variables leads to loss of information. Moreover, dichotomizing may add errors of discreteness to the measurement error in the original scales (Cohen, 1983). In classical test theory, (a) observed score variance is composed of true score variance plus error variance; (b) reliability is defined as the proportion of true score in observed score variance; and (c) the square root of reliability represents the upper possible value of validity that a measure can achieve (Gulliksen, 1950; Lord & Novick, 1968). Consequently, adding error to measurement scores will result in scores with lower reliability and thus with lower validity. Following Cohen's (1983) analyses, it can be assumed that

dichotomous scoring is unlikely to be the optimal format for the BIDR, and that continuous scoring is likely to be the superior alternative.

Unfortunately, there is a dearth of research comparing continuous and dichotomous scoring of BIDR scores in a systematic manner. Consequently, there is no data base from which any firm conclusions can be drawn about which scoring method is preferable. Moreover, only a few studies report analyses for both scoring methods (Booth-Kewley, Edwards, & Rosenfeld, 1992; Paulhus, 1994). However, the findings from these studies do seem to indicate that dichotomous scoring may indeed have some disadvantages compared to continuous scoring. First, the reliability of the scores from the dichotomous scoring procedure seems to be lower than that of the scores from the continuous scoring procedure. Paulhus (1994) reports that Cronbach's alphas of continuous BIDR scores typically are in the range of .70 to .82 for SDE and .80 to .86 for IM. In comparison, Cronbach's alphas of dichotomous scores typically are in the range of .65 to .75 for SDE and .75 to .80 for IM. Moreover, he warns that, for some samples, alphas can be more extreme, and may even slip below .65 (Paulhus, 1994, p. 8). Booth-Kewley et al. (1992) found that scores from the dichotomous scoring method were less sensitive to instructional variations than those from the continuous scoring method. Only when the continuous scoring method was applied did they find a significant main effect for anonymity level with higher scores under non-anonymous conditions. When the dichotomous scoring method was applied, neither SDE nor IM scores responded to experimental variations of anonymity

The aim of the present article is to provide a first comprehensive and systematic comparison of the reliability of scores from the two different BIDR scoring methods. To this end, we inspected differences between the two scoring methods with respect to (a) convergent correlations with other measures of social desirability, (b) sensitivity to

instructional variations, and (c) relationships with personality. Three studies are presented. In Study 1, the aim was to examine the two BIDR scoring methods with respect to concurrent correlations with two other measures of social desirability, namely the Mummendey-Eifler Scale (Mummendey & Eifler, 1993) and the Social Desirability Scale-17 (Stöber, 1999, in press). Whereas the Mummendey-Eifler Scale mainly captures self-deceptive enhancement, the Social Desirability Scale-17 primarily taps impression management (Stöber, in press). The goal of Study 2 was to examine the two scoring methods with respect to their sensitivity towards instructional variations. To this end, the BIDR was administered under standard instructions and under fake instructions, with participants instructed either to fake a good impression (fake good) or to fake a bad impression (fake bad), and differences between the two scoring methods were then compared. Finally, the objective of Study 3, was to compare the two scoring methods with respect to their correlations with measures of the five-factor model of personality. To this end, measures from Costa and McCrae's (1992) five-factor model were applied, and correlations with neuroticism, extraversion, openness, agreeableness, and conscientiousness were analyzed.

Study 1

Method

Participants and Procedure

A sample of N = 101 students was recruited at the Martin Luther University of Halle-Wittenberg, Germany. Of these, 79 were female and 20 male (two participants did not indicate their gender). Mean age was 22.2 years (SD = 3.3; range = 19-40; four participants did not indicate their age). Respondents completed a set of questionnaires that also included the three measures described below. Participants volunteered in exchange for two hours of extra course credit or a lottery ticket for a chance to win 100 German marks.

Measures

Balanced Inventory of Desirable Responding (BIDR). The BIDR Version 6 (Paulhus, 1994) contains 40 items; 20 items capture self-deceptive enhancement (SDE) (e.g., "I always know why I like things"; "It would be hard for me to break any of my bad habits," negatively keyed), and 20 items capture impression management (IM) (e.g., "When I hear people talking privately, I avoid listening"; "I sometimes drive faster than the speed limit," negatively keyed). In the present study, the German translation prepared by Musch (1999) was used. BIDR items were presented with a 7-point answer scale ranging from "Not true" (1) to "Very true" (7), because this is the answer format suggested by Paulhus (1994) and used by most authors. For each BIDR subscale, two scores were computed: Continuous scores were computed by reversing negatively keyed items and then summing answers across items. Dichotomous scores were computed by reversing negatively keyed items, awarding one point for each "6" or "7" response, and then summing points across items.

Mummendey-Eifler Scale (MES). The Mummendey-Eifler Scale of Social Desirability (Mummendey & Eifler, 1993) is a measure of social desirability consisting of items from the Trier Personality Inventory (TPI; P. Becker, 1989). It was constructed using the method of instructional variation. A student sample responded to selected items from the TPI subscales on Mental Health, Behavior Control, Autonomy, and Expansiveness, first under standard instructions and then under social-desirability provoking instructions. The items that showed the largest mean differences between conditions were selected for inclusion in the new scale. With the removal of redundant items, the scale was reduced to 12 items (e.g., "I am in good physical and mental condition" or "There are times when I cannot stand myself," reverse keyed). As the Mummendey-Eifler Scale is intended for use with variable answer formats depending on the research question (Hans Mummendey, personal e-

mail communication, June 4, 1999), the scale was administered with the classical dichotomous answer format of "True" (1) and "False" (0). With a Cronbach's alpha of .74, scores showed satisfactory reliability.

Social Desirability Scale-17 (SDS-17). The Social Desirability Scale-17 (Stöber, 1999, in press) is a measure of social desirability constructed in the style of the Marlowe-Crowne Scale (Crowne & Marlowe, 1960), but including new items with formulations and contents that are more up to date and thus correspond more closely to today's beliefs about socially desirable behaviors (e.g., " I always stay friendly and courteous with other people, even when I am stressed out"; "I sometimes litter," negatively keyed). Originally, the SDS-17 contained 17 items, thus its name (Stöber, 1999). Further validation studies, however, showed that one item on drug use consistently showed item-total correlations near zero, so the scale was reduced to 16 items (Stöber, in press). Like the Marlowe-Crowne Scale, the SDS-17 is presented with a dichotomous answer format of "True" (1) and "False" (0). With a Cronbach's alpha of .75, scores showed satisfactory reliability.

Results

Table 1 presents the means, standard deviations, Cronbach's alphas, and intercorrelations for the scores resulting from the two scoring methods. In line with previous findings (Paulhus, 1994), the scores from continuous scoring (continuous scores) yielded higher Cronbach's alphas than those from dichotomous scoring (dichotomous scores). To investigate whether these differences were reliable, the test of the equality of two Cronbach's alphas developed by Feldt (1980) was administered following the formula and procedures described by Charter and Feldt (1996, p. 767). For both BIDR subscales, results indicated that the continuous scores yielded significantly higher Cronbach's alphas than dichotomous scores: for SDE, $t(99) = 2.23$, and for IM, $t(99) = 1.73$, both $ps < .05$.

Next, the convergent correlations of continuous and dichotomous scores were examined (Table 2). In line with expectations, the two BIDR subscales showed a differentiated pattern of correlations. Both SDE scores showed substantial correlations with the Mummendey-Eifler Scale, but not with the Social Desirability Scale-17; in contrast, both IM scores showed substantial correlations with the Social Desirability Scale-17, but not with the Mummendey-Eifler Scale. However, the convergent correlations were more pronounced for the continuous scores than for the dichotomous scores. Statistical comparison of the correlations (Meng, Rosenthal, & Rubin, 1992) showed that the correlation with the Mummendey-Eifler Scale was significantly higher for the continuous SDE scores than for the dichotomous SDE scores. In comparison, the difference between the correlation with the Social Desirability Scale-17 of the continuous IM scores on the one hand and the dichotomous IM scores on the other only approached standard levels of significance.

Discussion

In sum, the findings of Study 1 indicate that BIDR scores derived from continuous scoring (continuous scores) differed from those derived from dichotomous scoring (dichotomous scores) with respect to both reliability and convergent validity. First, the internal consistency of the continuous scores was significantly higher than that of the dichotomous scores. In fact, the Cronbach's alpha of the dichotomous SDE scores fell below the .60 value that is usually considered the lower bound acceptable for research purposes (Carmines & Zeller, 1979). Second, the convergent correlations of the continuous scores with two other measures of social desirability were significantly higher than those of the dichotomous scores. The latter finding may be particularly noteworthy because the two other measures of social desirability contained a dichotomous answer format (true/false), thus capturing only extreme responses. The dichotomous scoring procedure was also introduced

to the BIDR with the aim of capturing only extreme responses. Despite this, the continuous BIDR scores showed higher correlations than the dichotomous scores.

However, convergent correlations with other measures of the same construct are only one way of judging the validity of social desirability scales. Another important aspect is how sensitive social desirability scales are to instructional variations. Research in social desirability has shown that particularly instructions to make a favorable (or unfavorable) impression have a decisive effect on participants' social desirability scores (e.g., Paulhus, Bruce, & Trapnell, 1995; Stöber, in press). Consequently, the aim of Study 2 was to compare the two scoring methods with respect to their sensitivity towards instructional variations. For this, the BIDR was administered under standard instructions and under fake instructions, with participants instructed either to fake a good impression (fake good) or to fake a bad impression (fake bad). In line with the findings reported by Paulhus et al. (1995, p. 103, Figure 2), we expected IM scores to show substantial increases under fake-good instructions and substantial decreases under fake-bad instructions relative to standard instructions. Moreover, we expected SDE scores to show only minor increases under fake-good instructions and no effects under fake-bad instructions.

Study 2

Method

Participants

A sample of N = 55 students was recruited at two high schools in Duisburg, Germany. Of these, 43 were female and 12 male. To obtain a more diverse sample and thus increase the generalizability of the findings, the sample included both younger students attending a standard high school (aged between 16 and 19 years) and older adults attending a night school.[1] Mean age was 20.4 years (SD = 7.5; range = 16-57). All participants

volunteered to take part in this experiment without any form of compensation, either financial or otherwise.

Measures and Procedure

All participants completed a set of questionnaires including the BIDR. As in Study 1, Version 6 of the BIDR (Paulhus, 1994; German translation: Musch, 1999) was applied. The BIDR was included twice: first under standard instructions (placed at the beginning of the set of questionnaires) and then again under fake-impression instructions (placed at the end of the set). There were two fake-impression instructions, to which participants were randomly allocated. One half of the participants (n = 28) received fake-good instructions; the other half (n = 27) received fake-bad instructions. The fake-good instructions read: "Now you will see some questions that you have answered before. This time, however, we would like you to imagine a situation in which you want to make as good an impression as possible, for example, a job-application situation. Therefore, please answer all question in such a way as to make as good an impression as possible." In contrast, the fake-bad instructions read: "Now you will see some questions that you have answered before. This time, however, we would like you to imagine a situation in which you want to make as bad an impression as possible. Therefore, please answer all question in such a way as to make as bad an impression as possible."

Results

Table 3 presents the means, standard deviations, Cronbach's alphas, and intercorrelations for the scores resulting from the two scoring methods under standard instructions. As in Study 1, the scores from continuous scoring (continuous scores) yielded higher Cronbach's alphas than those from dichotomous scoring (dichotomous scores). When Feldt's (1980) tests were computed to test the significance of these differences, the

continuous IM scores yielded a significantly higher Cronbach's alpha reliability coefficient than the dichotomous IM scoring procedure, $t(53) = 2.04$, $p < .05$. For SDE, the difference between the Cronbach's alphas of the two scores was not significant, $t(53) = 0.57$, ns.

Next, the effects of the instructional variations on the different scoring methods were examined, separately for each BIDR subscale (Figure 1). First, the effects for the scores from continuous scoring were examined (see top part of Figure 1). For the group that received fake-good instructions following standard instructions, results showed highly significant effects for both SDE and IM scores. Following fake-good instructions, SDE scores displayed an increase of $M = 14.04$ scale points, $SE = 2.34$, $t(27) = 5.99$, $p < .001$; and IM scores displayed an increase of $M = 30.96$ scale points, $SE = 5.90$, $t(27) = 5.50$, $p < .001$. For the group that received fake-bad instructions following standard instructions, results showed a highly significant effect for IM scores, but not for SDE scores. Following fake-bad instructions, SDE scores displayed a nonsignificant increase of $M = 6.37$ scale points, $SE = 4.45$, $t(26) = 1.43$, ns; whereas, in line with our expectations, IM scores displayed a decrease of $M = –21.30$ scale points, $SE = 5.08$, $t(26) = –4.19$, $p < .001$. Thus, for both BIDR subscales, the continuous scores showed effects in the expected direction.

Second, the effects for the scores from dichotomous scoring were examined (see bottom part of Figure 1). For the group that received fake-good instructions following standard instructions, results again showed highly significant effects for both SDE and IM scores. Following fake-good instructions, dichotomous SDE scores displayed an increase of $M = 4.04$ scale points, $SE = 0.63$, $t(27) = 6.40$, $p < .001$, and dichotomous IM scores displayed an increase of $M = 6.39$ scale points, $SE = 1.17$, $t(27) = 5.25$, $p < .001$. Thus, under fake-good instructions, dichotomous scores behaved exactly like continuous scores. Under fake-bad instructions, however, dichotomous scores behaved differently. Contrary to

expectations, dichotomous SDE scores displayed a substantial increase of M = 4.63 scale points, SE = 0.73. Moreover, this effect was highly significant, t(26) = 6.36, p < .001. In contrast, dichotomous IM scores displayed only a small decrease of M = –1.15 scale points, SE = 0.81. This effect was not significant, t(26) = –1.42, ns. Thus, under fake-bad instructions, dichotomous SDE scores showed an unexpected effect, while dichotomous IM scores failed to show the expected effect.

Discussion

In sum, the findings of Study 2 indicate that the BIDR scores derived from continuous scoring (continuous scores) differed from those derived from dichotomous scoring (dichotomous scores) with respect to both reliability and sensitivity to instructional variations. First, similarly to Study 1, the internal consistency of the continuous scores was again higher than that of the dichotomous scores. Differently from Study 1, however, the difference in Cronbach's alpha was significant only for IM scores, and not for SDE scores. Second, with respect to instructional variations, only the continuous scores behaved in line with expectations under both fake-good and fake-bad instructions. In contrast, the dichotomous scores displayed the expected effects only under fake-good instructions. Under fake-bad instructions, dichotomous IM scores failed to show significantly lower values, while dichotomous SDE scores showed significantly higher values than under standard instructions.

The failure to find a significant difference between the Cronbach's alphas of continuous and dichotomous SDE scores may simply be attributed to the low statistical power associated with the small sample size. The unexpected findings for the dichotomous scores under fake-bad instructions require some discussion, however. A possible explanation would be that participants did not understand the fake-bad instructions. As there was no

manipulation check, this might be a plausible explanation if it were not for fact that these instructions did result in the expected effect on the continuous IM scores (i.e., a substantial decrease in impression management). Consequently, it may be more plausible to assume that participants did understand the fake-bad instructions, and replied accordingly, but that they still avoided extremely low IM responses. Consequently, because it was less sensitive to small distortions in answer behavior, the dichotomous scoring procedure was unable to detect what the continuous scoring procedure had identified as a significant effect of the fake-bad instruction.

Participants chose extremely high SDE responses in the fake-bad condition. This seems to suggest that participants assume that demonstrating excessive levels of self-deception (i.e., demonstrating unrealistic beliefs that one is in total control of oneself) would make a bad impression, either because it indicates a lack of knowledge about the world and oneself, or because it presents a case of easy-to-detect bolstering, with the respondent claiming to have unrealistically good self-regulatory skills.

This interpretation may be supported by the findings of Paulhus et al. (1995), who compared honest instructions with fake-bad ("fake bad without arousing suspicion") and fake-worst ("fake the worst possible candidate") instructions using dichotomous BIDR scores. Results showed identical SDE scores for honest and fake-bad instructions, whereas fake-worst instructions showed significantly higher SDE scores than honest instructions. Thus, our findings regarding dichotomous SDE scores under fake-bad instructions are in line with Paulhus et al.'s (1995) findings for dichotomous SDE scores under fake-worst instructions. Nevertheless, they are not in line with common conceptions of self-deceptive enhancement in test responses (e.g., Paulhus, 1986).

In contrast, both continuous BIDR scores behaved according to expectations,

showing higher convergence with previous findings under all instructional variations. Consequently, we expected that continuous BIDR scores would also show higher convergence with previous findings on correlations with measures of the Big Five personality traits.

Study 3

Method

Participants and Procedure

A sample of N = 166 students, who had not participated in Study 1, was recruited at the Martin Luther University of Halle-Wittenberg. Of these, 88 were female and 75 male (three participants did not indicate their gender). Mean age was 23.6 years (SD = 3.4; range = 18-41; five participants did not indicate their age). Respondents completed a comprehensive questionnaire battery that also included the measures described below. Participants volunteered in exchange for a lottery ticket for a chance to win 100 German marks.

Measures

Balanced Inventory of Desirable Responding (BIDR). Because of the large number of questionnaires that was included in the battery, the short form of the BIDR (Musch, Brockhaus, & Bröder, 2001) was administered to reduce the load on participants. Musch et al. constructed the BIDR short form using Musch's (1999) translation of the BIDR Version 6 (Paulhus, 1994), computing factor analyses, and selecting those 10 items with the highest loadings on the two factors representing self-deceptive enhancement and impression management, respectively, for each BIDR dimension. Consequently, the short form of the BIDR contains 20 items, 10 of which items capture self-deceptive enhancement (SDE) and 10 of which tap impression management (IM).[2] Following Paulhus' (1994) recommendation,

items were again presented with a 7-point answer scale from "Not true" (1) to "Very true" (7).

Big Five personality traits. To measure the Big Five personality traits, the NEO Five Factor Inventory (NEO-FFI; Costa & McCrae, 1992; German version: Borkenau & Ostendorf, 1993) was included. The NEO-FFI consists of five scales that capture individual differences in neuroticism, extraversion, openness, agreeableness, and conscientiousness. Each scale comprises 12 items. Items are answered on a 5-point rating scale from "Strongly disagree" (0) to "Strongly agree" (4). With Cronbach's alphas from .69 to .88, all scales displayed satisfactory reliabilities.

Results and Discussion

Table 4 presents the means, standard deviations, Cronbach's alphas, and intercorrelations for the scores resulting from the two scoring methods. As in the two preceding studies, scores from continuous scoring (continuous scores) displayed higher Cronbach's alphas than scores from dichotomous scoring (dichotomous scores). Moreover, like in Study 1, the differences between the respective Cronbach's alphas were significant for both scales: for SDE, $t(164) = 4.05$, $p < .001$, and for IM, $t(164) = 3.18$, $p < .01$.

Next, the correlations of scores with the measures of the Big Five personality traits were examined (Table 5). First, the correlations of SDE scores were inspected. In line with previous findings suggesting that high SDE scorers usually are well-adjusted individuals (e.g., Paulhus, 1994), SDE scores showed substantial negative correlations with neuroticism and substantial positive correlations with conscientiousness. Moreover, there was a small positive correlation with extraversion, but only for the continuous SDE scores. In addition, dichotomous SDE scores showed a small negative correlation with agreeableness. When the correlations of the continuous SDE scores were compared with those of the dichotomous

SDE scores following the procedures of Meng et al. (1992), results indicated that the correlations of the continuous SDE scores were significantly higher than those of the dichotomous SDE scores for all traits for which correlations with SDE were expected (i.e., neuroticism, extraversion, and conscientiousness).

Second, the correlations of IM scores were inspected. In line with previous findings suggesting that high IM scorers are socially conventional and cautious individuals (e.g. Paulhus, 1994), IM scores showed substantial positive correlations with agreeableness and conscientiousness. In addition, IM scores showed negative correlations with openness and extraversion. The correlation with extraversion was only significant for continuous IM scores, however. When the differences between the correlations of continuous and dichotomous IM scores were tested for significance, significant differences emerged only for those traits for which correlations with IM were not expected (i.e., neuroticism, openness, and--marginally--extraversion). In contrast, the correlations of continuous and dichotomous IM scores did not differ for those traits for which correlations with IM were expected (i.e., agreeableness and conscientiousness).

In sum, the findings of Study 3 again demonstrate that the BIDR scores derived from continuous scoring are significantly more reliable (Cronbach's alpha) than those derived from dichotomous scoring. Moreover, results indicate that continuous SDE scores display significantly higher correlations with the Big Five personality traits for which previous research has found correlations (i.e., neuroticism, extraversion, and conscientiousness) than do dichotomous SDE scores. For IM scores, however, continuous scores do not display higher correlations with those personality traits for which previous research has found substantial correlations (i.e., agreeableness and conscientiousness) than do dichotomous scores.

General Discussion

Three studies were presented to investigate differences between the two different scoring methods suggested by Paulhus (1994) for the two subscales of the Balanced Inventory of Desirable Responding (BIDR). With continuous scoring, scores are computed by summing all the answers on the Likert-style answer scale across items. With dichotomous scoring, only extreme answers are counted. Suggesting that only extreme answers indicate socially desirable responding, Paulhus (1994) recommended the dichotomous scoring method as the optimal strategy to obtain accurate measures of self-deceptive enhancement (SDE) and impression management (IM) on the BIDR subscales. In contrast, the present studies show that dichotomous scoring may be suboptimal. In fact, our findings show that the BIDR scores derived from continuous scoring (continuous scores) had some clear advantages over the BIDR scores derived from dichotomous scoring (dichotomous scores). First, across studies, continuous scores showed higher Cronbach's alphas than dichotomous scores, indicating higher reliability of measurement with continuous scoring. Second, continuous scores showed higher convergent correlations with other measures of social desirability, even though these measures used a dichotomous true-false answer format. Third, only the continuous scores showed the expected effects of instructional variations: substantially higher IM scores (and somewhat higher SDE scores) after fake-good instructions, and lower IM scores after fake-bad instructions. In contrast, dichotomous scores showed the expected effects only for fake-good instructions, but did not yield the expected effects under fake-bad instructions. Finally, continuous SDE scores displayed higher correlations with traits for which substantial associations with SDE were expected than did dichotomous scores.

In sum, the present findings indicate that when continuous scoring is used for the

BIDR subscales, the resulting scores will show higher reliability, higher sensitivity to instructional variations, and higher convergent correlations than those derived from dichotomous scoring. Thus, based on these findings, we suggest that continuous scoring, and not dichotomous scoring, is used when applying the BIDR.

However, there are some potential limitations that may require further research before stronger recommendations can be made. First, Cronbach's alpha captures the internal consistency (homogeneity) of a scale and represents only a lower-bound estimate of reliability (Crocker & Algina, 1986). Social desirability scales, however, usually consist of items with quite heterogeneous content. In this case, test-retest correlations may be the more appropriate statistics and yield higher estimates of reliability than Cronbach's alpha, as could be demonstrated for the Social Desirability Scale-17 (Stöber, 1999). For the BIDR subscales, however, reported test-retest correlations are in the same order of magnitude as the Cronbach's alphas (Paulhus, 1994) or lower (Paulhus, 1991). Therefore, it is doubtful that investigating reliability with test-retest correlations would have led to substantially different results than those reported in the present article.

Second, the samples for Studies 1 and 2 contained only few male participants. Therefore, these findings may be limited to female participants, even though we did not find any evidence for systematic gender effects in any of the three studies. Moreover, the present findings were obtained using a German translation of the BIDR Version 6, not the original inventory. Even though great care was taken in the translation of the original BIDR, using backward and forward translations and consulting native English speakers (Musch et al., 2001), literal translations of questionnaires may sometimes fail to capture the same psychological content as the original (Van de Vijver & Hambleton, 1996). So far, however, all studies implementing the German translation of the BIDR have obtained findings

comparable to those of studies that applied the original BIDR (e.g., Musch et al., 2001; Stöber, in press). Therefore, we can be fairly confident that the German BIDR is an adequate translation of the original BIDR, both literally and psychologically.

Finally, it is possible that the present findings, which were obtained using the 7-point answer format, are not generalizable to applications of the BIDR using the 5-point answer format. Consequently, one may argue that dichotomous scoring of the BIDR shows clear disadvantages only in combination with the 7-point answer format. Like Booth-Kewley et al. (1992), our studies used the BIDR with the 7-point answer format and found that only the continuous scores showed the expected effects of instructional variations. In contrast, using the 5-point answer format with dichotomous scoring, Paulhus et al. (1995) found that all scores showed the expected effects of instructional variations. However, as Paulhus et al. did not compare dichotomous scores with continuous scores, and as neither we nor Booth-Kewley et al. (1992) investigated scoring differences with the 5-point answer format, this issue remains unresolved and requires further research. Still, for the 7-point answer format, our findings do indicate that continuous scoring may be preferable to dichotomous scoring when calculating subscale scores for the Balanced Inventory of Desirable Responding.

References

Becker, G., & Cherny, S. S. (1992). A five-factor nuclear model of socially desirable responding. Social Behavior and Personality, 20, 163-192.

Becker, P. (1989). Der Trierer Persönlichkeitsfragebogen: TPF [The Trier Personality Inventory: TPI]. Göttingen, Germany: Hogrefe.

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

Borkenau, P., & Ostendorf, F. (1993). NEO-Fünf-Faktoren Inventar (NEO-FFI) [NEO Five Factor Inventory (NEO-FFI)]. Göttingen, Germany: Hogrefe.

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Beverly Hills, CA: Sage.

Charter, R. A., & Feldt, L. S. (1996). Testing the equality of two alpha coefficients. Perceptual and Motor Skills, 82, 763-768.

Cohen, J. (1983). The cost of dichotomization. Applied Psychological Measurement, 7, 249-253.

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

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

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

Davies, M. F., French, C. C., & Keogh, E. (1998). Self-deceptive enhancement and impression management correlates of EPQ-R dimensions. Journal of Psychology, 132, 401-406.

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

Eysenck, H. J., & Eysenck, S. B. G. (1964). Manual of the Eysenck Personality Inventory. London: University of London Press.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Psychometrika, 45, 99-105.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. Psychological Bulletin, 111, 172-175.

Meston, C. M., Heiman, J. R., Trapnell, P. D., & Paulhus, D. L. (1998). Socially desirable responding and sexuality self-reports. Journal of Sex Research, 35, 148-157.

Mummendey, H. D., & Eifler, S. (1993). Eine neue Skala zur Messung Sozialer Erwünschtheit [A new scale for the measurement of social desirability] (Bielefelder Arbeiten zur Sozialpsychologie Nr. 167). Bielefeld, Germany: University of Bielefeld.

Musch, J. (1999). German version of the Balanced Inventory of Desirable Responding (BIDR version 6). Unpublished manuscript, University of Bonn, Germany.

Musch, J., Brockhaus, R., & Bröder, A. (2001). Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit [An inventory to assess two factors of social desirability]. Manuscript submitted for publication, University of Bonn, Germany.

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

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires (pp. 143-165). Berlin, Germany: Springer.

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.

Paulhus, D. L. (1994). Balanced Inventory of Desirable Responding: Reference manual for BIDR Version 6. Unpublished manuscript, University of British Columbia, Vancouver, Canada.

Paulhus, D. L. (1998). The Balanced Inventory of Desirable Responding. Toronto, Canada: Multi-Health Systems.

Paulhus, D. L. (in press). Socially desirable responding: Evolution of a construct. In H. Braun, D. N. Jackson, & D. E. Wiley (Eds.), The role of constructs in psychological and educational measurement. Hillsdale, NJ: Erlbaum.

Paulhus, D. L., Bruce, M. D., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. Personality and Social Psychology Bulletin, 21, 100-108.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. Journal of Personality and Social Psychology, 60, 307-317.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the validity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. Journal of Personality and Social Psychology, 78, 582-593.

Sackeim, H. A., & Gur, R. C. (1978). Self-deception, other-deception, and consciousness. In G. E. Schwartz & D. Shapiro (Eds.), Consciousness and self-regulation: Advances in research (Vol. 2, pp. 139-197). New York: Plenum Press.

Stöber, J. (1999). Die Soziale-Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zu Reliabilität und Validität [The Social Desirability Scale-17 (SDS-17): Development and first results on reliability and validity]. Diagnostica, 45, 173-177.

Stöber, J. (in press). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. European Journal of Psychological Assessment.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. European Psychologist, 1, 89-99.

Author Note

For the present article, data from studies originally presented in Musch et al. (2001) and Stöber (in press) were reanalyzed. We are grateful to Robbi Brockhaus for her help in collecting the data for Study 2. Moreover, we would like to express our gratitude to Claudia Dalbert and two anonymous reviewers for helpful comments on an earlier version of this article.

All correspondence concerning this article should be addressed to Joachim Stöber, Martin Luther University of Halle-Wittenberg, Department of Educational Psychology, Franckesche Stiftungen, Haus 5, D-06099 Halle (Saale), Germany. E-mail: stoeber@paedagogik.uni-halle.de.

Footnotes

[1]A night school is an open school with evening classes for people who, for various reasons, did not acquire a high-school diploma (in German "Abitur") in their younger days.

[2]The BIDR short form consists of the items numbered 1, 4, 5, 10, 12, 15, 16, 17, 18, and 20 (SDE) and 21, 23, 24, 25, 29, 30, 33, 35, 36, and 37 (IM) from the BIDR Version 6.

Table 1

Study 1: Means, Standard Deviations, Cronbach's Alphas, and Intercorrelations

BIDR subscale	Scoring	M	SD	α	Correlation 1	2	3
1. Self-Deceptive Enhancement (SDE)	Continuous	77.21	13.11	.69	--		
2. Self-Deceptive Enhancement (SDE)	Dichotomous	4.88	2.74	.59	.79***	--	
3. Impression Management (IM)	Continuous	69.69	15.98	.73	.35***	.21*	--
4. Impression Management (IM)	Dichotomous	4.99	3.10	.68	.39***	.39***	.84***

Note. N = 101. BIDR = Balanced Inventory of Desirable Responding. α = Cronbach's alpha.

*p < .05. ***p < .001.

Table 2

Study 1: Convergent Correlations with Two Measures of Social Desirability

			BIDR subscale					
			Self-Deceptive Enhancement (SDE)			Impression Management (IM)		
			Scoring			Scoring		
Measure	M	SD	Continuous	Dichotomous	z(diff)	Continuous	Dichotomous	z(diff)
Mummendey-Eifler Scale	7.92	2.62	.61***	.45***	2.96**	.21*	.22*	–0.02
Social Desirability Scale-17	8.53	3.45	.20*	.16	0.50	.46***	.37***	1.91^{+}

Note. N = 101. BIDR = Balanced Inventory of Desirable Responding. z(diff) = z value of difference between correlations (Meng et al., 1992).

$^{+}p < .06$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table 3

Study 2: Means, Standard Deviations, Cronbach's Alphas, and Intercorrelations (Under Standard Instructions)

					Correlation		
BIDR subscale	Scoring	M	SD	α	1	2	3
1. Self-Deceptive Enhancement (SDE)	Continuous	74.31	10.83	.55	--		
2. Self-Deceptive Enhancement (SDE)	Dichotomous	3.87	2.32	.49	.69***	--	
3. Impression Management (IM)	Continuous	66.31	14.72	.72	.20	.01	--
4. Impression Management (IM)	Dichotomous	4.40	2.61	.61	.26	.19	.82***

Note. N = 55. BIDR = Balanced Inventory of Desirable Responding. α = Cronbach's alpha.

***p < .001.

Table 4

Study 3: Means, Standard Deviations, Cronbach's Alphas, and Intercorrelations

					Correlation		
BIDR subscale	Scoring	M	SD	α	1	2	3
1. Self-Deceptive Enhancement (SDE)	Continuous	40.30	7.34	.66	--		
2. Self-Deceptive Enhancement (SDE)	Dichotomous	2.03	1.64	.46	.67***	--	
3. Impression Management (IM)	Continuous	32.06	8.62	.67	.12	.05	--
4. Impression Management (IM)	Dichotomous	1.81	1.65	.55	.09	.15	.79***

Note. N = 166. BIDR = Balanced Inventory of Desirable Responding (short form). α = Cronbach's alpha.

***p < .001.

Table 5

Study 3: Correlations with Measures of the Big Five Personality Traits

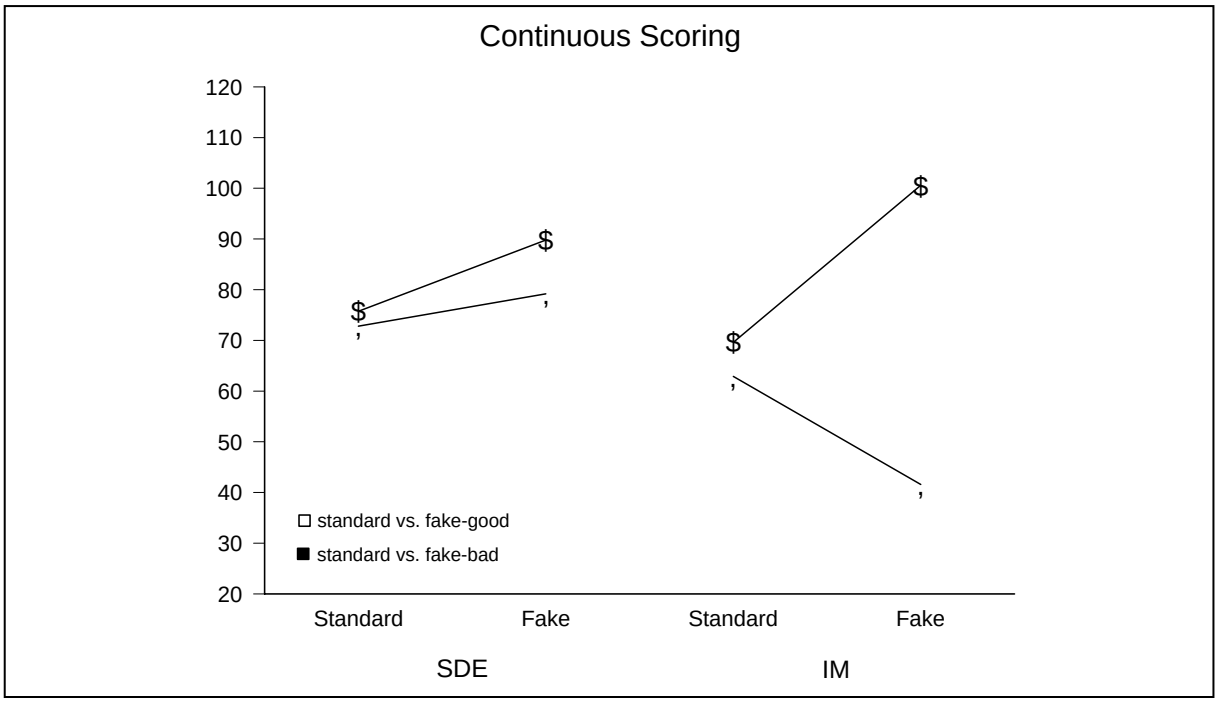
	BIDR subscale					
	Self-Deceptive Enhancement (SDE)			Impression Management (IM)		
	Scoring			Scoring		
Measure	Continuous	Dichotomous	z(diff)	Continuous	Dichotomous	z(diff)
Neuroticism	–.51***	–.31***	–3.54***	.04	–.06	2.11*
Extraversion	.17*	.03	2.21*	–.13*	–.04	–1.92^{+}
Openness	–.18*	–.11	–1.13	–.28***	–.18*	–1.99*
Agreeableness	–.12	–.17*	0.82	.30***	.35***	–1.09
Conscientiousness	.41***	.27***	2.28*	.39***	.33***	1.30

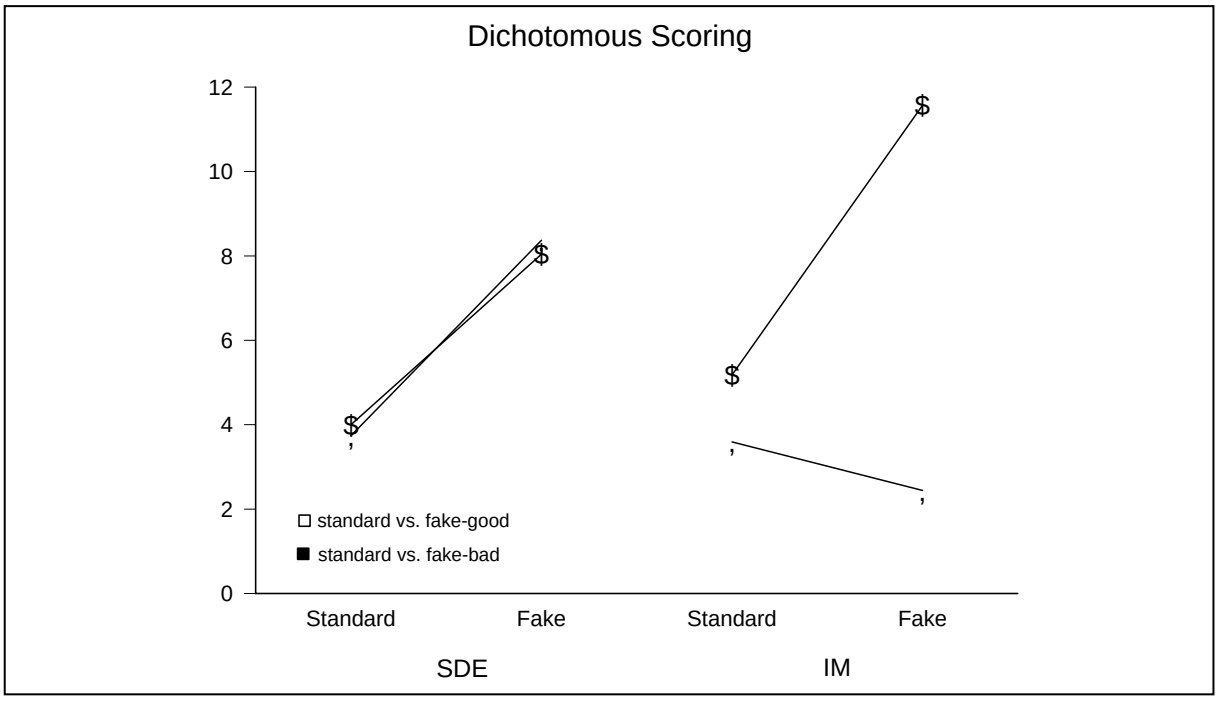
Note. N = 166. BIDR = Balanced Inventory of Desirable Responding (short form). z(diff) = z value of difference between correlations (Meng et al., 1992).

$^{+}p < .06$. *$p < .05$. ***$p < .001$.

Figures

Figure 1. BIDR Scores from Continuous and Dichotomous Scoring under Standard and Fake-Impression Instructions. SDE = Self-Deceptive Enhancement, IM = Impression Management.


Continuous Scoring
120
110
100
90
80
70
60
50
40
30
20
standard vs. fake-good
standard vs. fake-bad
Standard
Fake
Standard
Fake
SDE
IM


Dichotomous Scoring
12
10
8
6
4
2
0
standard vs. fake-good
standard vs. fake-bad
Standard
Fake
Standard
Fake
SDE
IM