Journal of Research in Personality 61 (2016) 80–86

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

The Marlowe–Crowne Social Desirability Scale outperforms the BIDR Impression Management Scale for identifying fakers ☆

Christine E. Lambert *, Spencer A. Arbuckle, Ronald R. Holden

Queen's University, Kingston, Ontario, Canada

ARTICLE INFO

Article history:
Received 12 August 2015
Revised 12 February 2016
Accepted 16 February 2016
Available online 17 February 2016

Keywords:
Faking
Socially desirable responding
Self-report
Faking detection
Marlowe–Crowne Social Desirability Scale
Balanced Inventory of Desirable Responding

ABSTRACT

Self-report personality tests are used widely, but it is not uncommon for an individual's scale score to be invalid due to Socially Desirable Responding (SDR): answering to be viewed favourably. Various indices exist to detect SDR (e.g., faking). The Marlowe–Crowne Social Desirability Scale (MCSDS) formerly was the most popular. The current gold standard is the Balanced Inventory of Desirable Responding (BIDR), considered more sensitive because its development incorporated newer theoretical and empirical understanding of SDR and more sophisticated multivariate techniques. We compare the efficacy of these measures with surprising results: the MCSDS consistently outperforms the BIDR in identifying fakers. This finding indicates that the MCSDS should be retained because it captures elements of faking more effectively than the modern scale.

1. Introduction

Self-report personality tests are used in many settings throughout society. They help individuals discover their vocational interests, assess the clinical status of forensic, counselling, and psychiatric patients, and evaluate the suitability of job applicants. However, although personality tests are developed to yield scores that are valid predictors of relevant criteria across large samples, it is not uncommon for a particular individual's scale score to be invalid because of faking (Butcher, Morfitt, Rouse, & Holden, 1997; Rosse, Stetcher, Miller, & Levin, 1998). Holden and Book (2012, p. 71) define faking as "intentional misrepresentation in self-report." Participants are likely to fake results in high-stakes situations in an attempt to increase their chances of attaining a desired outcome. They may "fake good" by exaggerating their positive characteristics on an integrity assessment for a job application, or "fake bad" by underperforming in an assessment of academic abilities in order to qualify for additional support (Holden, 2007; Viswesvaran & Ones, 1999). Faking good – the tendency to answer in a way that will be viewed favorably by others – has also been termed Socially Desirability Responding (SDR), although it may represent only one type of SDR. Faking bad has received less research attention than faking good, but is an equally important phenomenon.

Accordingly, detecting and preventing faking on self-report personality inventories has become a matter of theoretical and practical importance. In test development, many personality inventories include validity indices. The Minnesota Multiphasic Personality Inventory (MMPI-2), which is frequently used for screening applicants for jobs that have a direct effect on public safety or security, includes seven validity indices (Butcher et al., 2001). Other entire inventories have been developed to assess individuals' response styles, such as the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1998) and the Marlowe–Crowne Social Desirability Scale (Crowne & Marlowe, 1960).

The utility of social desirability measures in assessing and adequately detecting applicant faking has been a very contentious issue, with some researchers arguing passionately against their use (Burns & Christiansen, 2006; Griffith & Peterson, 2008). This argument is based on findings in some studies that when measures of social desirability are used as a proxy of faking behavior, faking does not appear to affect criterion-related validity. Others (e.g., Mueller-Hanson, Heggestad, & Thornton, 2003), however, have demonstrated that the effect of faking can be impactful depending on moderating factors such as selection ratio. Findings have also been interpreted as indicating that social desirability and faking can be distinct, but related, constructs (Holden & Book, 2012). However, despite varying perspectives, measures of social desirability as indicators of faking continue to be widely used both in research and clinical practice.

☆ This research was supported by the Social Sciences and Humanities Research Council of Canada.

* Corresponding author at: Department of Psychology, Queen's University, Kingston, Ontario K7L 3N6, Canada.
E-mail address: 41cel2@queensu.ca (C.E. Lambert).

http://dx.doi.org/10.1016/j.jrp.2016.02.004
0092-6566/

Until recently, the Marlowe–Crowne Social Desirability Scale (MCSDS), originally published in the 1960s, was the most popular measure of SDR. The MCSDS consists of 33 items that were selected to have socially desirable content and low probability of occurrence (sample item: "I never hesitate to go out of my way to help someone in trouble"). Participants respond to each item by indicating whether it is true or false. High scores indicate that a respondent is presenting him/herself in an unrealistically favorable manner. The MCSDS scale scores had an internal reliability coefficient alpha of .88 in a sample of undergraduate students, and high concurrent validity as established through correlations with the MMPI validity scales (Crowne & Marlowe, 1960). More recently, a study conducted with an adapted version of the scale yielded Cronbach's alpha levels of 0.63 in Kenya, 0.66 in Mozambique, 0.70 in Uganda, and 0.80 in Ethiopia (Vu, Tran, Pham, & Ahmed, 2011).

The Balanced Inventory of Desirable Responding (BIDR; also published as the Paulhus Deception Scales) measures an individual's tendency to give socially desirable responses on self-report inventories. It consists of 40 items, with forms that are either rated on 7-point scales (1 = Totally Disagree; 4 = Neutral; 7 = Totally Agree) or 5-point scales (1 = Not True; 5 = Very True). Regardless of the response form, items are scored dichotomously. The BIDR contains two scales: *Self-Deceptive Enhancement* (SDE), the tendency to unconsciously give unrealistically favorable self-descriptions; and *Impression Management* (IM), the tendency to dissimulate by giving unrealistically positive self-descriptions (Paulhus, 1998). SDE occurs at an unconscious level, and measures an individual's honest, though inaccurate, beliefs about him/herself. In contrast, IM measures a conscious effort to dissimulate or fake good. Higher scores indicate greater tendencies toward SDE and IM. A BIDR Total scale can also be scored and is the sum of the SDE and IM scales. The BIDR Total scale scores had a coefficient alpha of .83, with scale scores having reliabilities of .70 (SDE) and .81 (IM) in a college student sample (Paulhus, 1998). IM scale scores have high concurrent validity with the Marlowe–Crowne Social Desirability Scale (r = .63; Helmes & Holden, 2003).

The IM scale of the BIDR is presently the most widely used validity index in detecting SDR, in general, and respondent faking, in particular (Davis, Thake, & Weekes, 2012; Pauls & Crost, 2004). Currently, when researchers and clinicians seek to establish whether an individual may be misrepresenting himself or herself on a personality inventory, they are quite likely to assess this by administering the IM scale of the BIDR, along with the rest of their personality measures. This is largely because, unlike older validity measures such as the MCSDS, the construction of the BIDR in the 1990s was based on sophisticated multivariate test construction techniques that were either not developed or not readily accessible in previous decades. The BIDR has the added benefit of providing cut-off scores for invalidity detection. Participants with scores greater than 12 and 8 are designated as probably and may be faking good, respectively, whereas participants with scores less than 1 and 2 are designated as probably and may be faking bad, respectively (Paulhus, 1998, p. 10).

Given the former and current popularity of the MCSDS and IM scale, it is surprising that no studies have compared the two scales for their respective abilities to correctly identify dissimulating respondents (i.e., fakers). As such, the goal of the current research was to evaluate the relative merits of these two premier validity scales in the detection of fakers. Based on the IM scale having been developed using more recent test construction practices, including more advanced multivariate item selection procedures, it was hypothesized that:

Hypothesis. The IM scale will be more accurate than the MCSDS in detecting respondents who are faking.

2. Method

2.1. Participants

Undergraduate students from a midsize university were recruited to participate using an introductory psychology course subject pool and by posting flyers around campus. Participants in Studies 1 and 2 were compensated with either course credit or $15 for an hour of their time.

2.1.1. Study 1

Two hundred and ninety-four students were recruited to participate in Study 1. The data from one participant were lost due to a computer malfunction, resulting in responses from a total of 293 individuals (66 men, 227 women) being included in the analyses. The participants were between the ages of 17 and 24 years (M = 18.82, SD = 1.04).

2.1.2. Study 2

Three hundred undergraduate students (57 men, 243 women) participated in Study 2. The participants were between the ages of 17 and 28 years (M = 19.22, SD = 1.79).

2.1.3. Study 3

One hundred and sixteen undergraduate students (14 men, 102 women) participated in Study 3. Participants ranged between 18 and 22 years in age (M = 19.78, SD = 0.88). Each participant received $5 compensation for his/her participation.

2.2. Materials

Participants in all three studies completed the MCSDS (Crowne & Marlowe, 1960) and (Paulhus', 1998) IM scale. In Studies 1 and 2, the SDE scale was also administered. For the IM and SDE scales, a 7-point Likert-type rating scale was used for Studies 1 and 2 (1 = Totally Disagree; 4 = Neutral; 7 = Totally Agree), whereas a 5-point rating scale was used for the IM scale in Study 3 (1 = Not True; 3 = Neutral; 5 = Very True). Participants in Study 3 also completed the Holden Applicant Reliability Measure (HARM; Holden, 2011). The HARM is a 100-item true/false self-report inventory that uses content directly related to on-the-job behavior in assessing eight aspects (e.g., dishonesty, drug use) of employee counterproductivity. Items include "My safety on the job has been affected by my use of alcohol" on the Alcohol Use scale, and "I have called in sick to work when I've been perfectly healthy" on the Unauthorized Absenteeism scale. Internal consistency reliability and validity for HARM scale scores have been demonstrated for university students (coefficient alpha = .87; Holden, Starzyk, Edwards, Book, & Wasylkiw, 2003) and for unemployed persons actively seeking work (coefficient alpha = .95; Holden, 2000).

2.3. Procedure

2.3.1. Study 1 and Study 2

Each experimental session began with obtaining written informed consent. Participants were then asked to answer the MCSDS and the IM and SDE scales as if they were being screened for military induction under 1 of 3 conditions. Participants were randomly assigned to: (1) complete the measures under standard instructions; (2) fake answers to maximize their chances of being inducted (i.e., fake good); or (3) fake answers to minimize their chances of being inducted (i.e., fake bad). All participants were warned of the presence of validity checks to detect faking, were asked to do their best to avoid being detected, and were given an incentive to do so: for each 25 participants, a $50 prize was

awarded to the participant who followed instructions most effectively and therefore was farthest from activating the validity checks. Following completion of the questionnaires, participants were asked to indicate their degree of compliance to their condition-specific faking instructions. Responses were made on a 9-point Likert-type scale ranging from 1 (Definitely did not comply with my instructions) to 9 (Always complied with my instructions). This served as a manipulation check to ensure that individuals were indeed responding to the questionnaire items in a manner that was consistent and in accordance with their instructions.

2.3.2. Study 3

Following informed consent, participants were randomly assigned to either the control ($n = 60$) or the fake good ($n = 60$) condition. Participants in the control condition received standard instructions for all study materials, whereas those in the fake good condition were instructed to fake their answers on the MCSDS, the IM scale, and the HARM to maximize their chances of being hired for a sensitive government job involving the handling of money and confidential material. As in Studies 1 and 2, participants in both the control and experimental conditions were advised that the survey included features that were designed to detect faking, which they wanted to avoid or "fake out." However, although participants in Study 3 were warned of the presence of validity checks, they were offered no reward for avoiding them.

3. Results

3.1. Manipulation check

Participants in both Study 1 and Study 2 indicated that they did comply with instructions on a 9-point Likert scale ($M = 7.38$, $SD = 1.52$ for Study 1; $M = 6.60$, $SD = 1.74$ for Study 2). As expected with high compliance, scale distributions were negatively skewed with more than 92% of respondents scoring at or above the manipulation check scale midpoint. In Study 1, compliance with instructions, that were specific to faking condition, was not related to scores on the MCSDS, $r(291) = .00$, $p > .05$, or the IM scale, $r(291) = .07$, $p > .05$. Similarly, in Study 2, compliance scores were not related to scores on the MCSDS, $r(291) = -.04$, $p > .05$, or the IM scale, $r(298) = .00$, $p > .05$. These results suggest it was reasonable to assume that compliance with instructions was not confounded with faking. An independent-samples t-test for the means of the fake-good ($M = 4.56$, $SD = 5.36$, coefficient $\alpha = .71$) and honest ($M = 7.95$, $SD = 5.95$, coefficient $\alpha = .67$) groups on the HARM showed that the two were significantly different, $t(112) = 3.19$, $p = .002$, indicating that the faking manipulation was also effective in Study 3.[1]

3.2. Discriminant function analyses

For each study, separate discriminant function analyses evaluated the ability of each measure of faking – the MCSDS, the IM, SDE, and BIDR Total scales in Studies 1 and 2; only the MCSDS and the IM scale in Study 3 – to correctly classify individuals into experimentally-induced faking groups. Because the base rate for faking is unknown with estimates varying widely (Griffith & Converse, 2012) and because participants in our studies were randomly assigned with equal probability to instructed faking conditions, analyses assumed equal Bayesian prior probabilities.

In all three studies where used, each of the MCSDS, IM, SDE, and BIDR Total scales were significant predictors of faking condition in the individual discriminant function analyses (see Table 1). In Study 1, compared to a chance rate of 33.3%, cross-validated successful classification rates were 76.7%, 67.9%, 46.4%, and 63.1% for the MCSDS, IM, SDE, and BIDR Total scales, respectively. In Study 2, corresponding cross-validated correct classification rates of 76.7%, 58.7%, 46.0%, and 52.7%, respectively, replicated the Study 1 relative ordering of the scales in successfully identifying fakers, and again indicated the weakest performance for the SDE scale. Interestingly, the performance of the BIDR Total scale, which includes both the IM and SDE scales, was lower than for the IM scale alone, indicating that the SDE scale attenuates and does not have added value relative to using only the IM scale when employing the Balanced Inventory of Desirable Responding (Paulhus, 1998). In Study 3, where only the MCSDS and IM scale were evaluated and the chance rate of identification was 50%, each of the MCSDS and the IM scale successfully identified participant group membership at a rate of 77.6%.[2]

Because classification rates using the BIDR Total scale were less than those associated with using only the IM scale, follow-up analysis no longer included the SDE scale and sought to compare only the MCSDS and IM scale when considered simultaneously in discriminant function analyses.[3] For Study 1, the combination of MCSDS and the IM scale significantly differentiated among faking conditions, $\lambda = .29$, $\chi^2(4, N = 293) = 354.96$, $p < .001$, with a cross-validated correct classification rate of 76.5%. In Study 2, the MCSDS and IM scale, in combination, again significantly discriminated among faking conditions, $\lambda = .33$, $\chi^2(4, N = 300) = 325.93$, $p < .001$, with a cross-validated correct identification rate of 74.4%. For Study 3, the MCSDS IM combination differentiated significantly between faking conditions, $\lambda = .64$, $\chi^2(2, N = 116) = 50.18$, $p < .001$, with a cross-validated correct classification rate of 75.9%. For each of the three studies, Table 2 displays canonical loadings and standardized function coefficients. In each study, the primary differentiating function more strongly incorporated the MCSDS relative to the IM scale, whether examined using standardized weights or canonical loadings. These results support the conclusion that the MCSDS works just as well, if not better, at detecting faking alone than the IM scale.

3.3. Effect sizes

Measures of effect size provide a meaningful indication of the strength of the difference in group detection between the MCSDS and the IM scale. For each measure, in each study, we calculated the strength of the measure for detecting individuals faking good and individuals faking bad relative to those following standard instructions (see Table 3). Study 3 had no Fake Bad condition, so for this study effect sizes were computed only for individuals faking good relative to those following standard instructions.

As is clear from Table 3, all of the Cohen's d effect sizes vastly surpassed the 0.80 guideline for indicating a large effect (Cohen,

[1] In the Study 3 manipulation check, the coefficient alpha reliabilities for the HARM scale scores were not as high as those previously reported (Holden, 2000; Holden et al., 2003). Inspection of HARM scale score distributions indicated floor effects ($z_{skew} = 5.15$, $p < .01$, and $z_{skew} = 3.48$, $p < .01$, for faking good and honest groups, respectively), a result associated with the manipulation check comprising counterproductive job behaviors, something that a sample of first-year university students (predominantly women) does not manifest to any large degree. The presence of a significant effect having a Cohen's $d = .61$ confirms an effect whose size estimation is spuriously underestimated by the floor effects that produce a corresponding attenuation in reliability estimates.

[2] Analyses using continuous rather than dichotomous scoring of the BIDR items did not substantively affect IM scale findings but did slightly improve results for the SDE and BIDR Total scales. Regardless of BIDR item scoring variation, the MCSDS remained as the optimum scale for correct classification. Given that dichotomous BIDR item scoring is the in-practice and test-manual-recommended procedure, we report results based on dichotomous item scoring.

[3] This emphasis on the IM scale rather than the SDE scale or Total BIDR scale also reflects current practical usage whereby the IM scale is extracted (in administration or just in analysis) from the rest of the BIDR.

Table 1
Discriminant function analyses: classification by lie scale.

	Lie scale	Cross-validated % correctly classified[a]	Chance rate %	Wilks' λ	Test statistic	p
Study 1	MCSDS	76.1% (79.4%)	33.3	.30	$\chi^2(2, N = 293) = 349.06$	<.001
	IM	67.9% (69.0%)	33.3	.44	$\chi^2(2, N = 293) = 236.29$	<.001
	SDE	46.4% (47.3%)	33.3	.82	$\chi^2(2, N = 293) = 58.42$	<.001
	BIDR Total	63.1% (62.5%)	33.3	.51	$\chi^2(2, N = 293) = 197.18$	<.001
Study 2	MCSDS	76.7% (80.5%)	33.3	.34	$\chi^2(2, N = 300) = 324.33$	<.001
	IM	58.7% (60.3%)	33.3	.55	$\chi^2(2, N = 300) = 175.58$	<.001
	SDE	46.0% (49.4%)	33.3	.89	$\chi^2(2, N = 300) = 33.78$	<.001
	BIDR Total	52.7% (53.9%)	33.3	.65	$\chi^2(2, N = 300) = 127.23$	<.001
Study 3	MCSDS	77.6%	50	.65	$\chi^2(2, N = 116) = 48.29$	<.001
	IM	77.6%	50	.70	$\chi^2(2, N = 116) = 41.12$	<.001

Note. MCSDS = Marlowe–Crowne Social Desirability Scale; IM = Impression Management Scale; SDE = Self-Deceptive Enhancement Scale; BIDR Total = Sum of IM and SDE. For cross-validated % correctly classified, each case is classified based on the discriminant function derived from all cases other than that case.
[a] Parenthesized value is when participants (⩽8%) with scores below the compliance check scale midpoint are omitted.

Table 2
Discriminant function analysis for identifying faking condition using the Marlowe–Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding Impression Management Scale simultaneously.

Discriminant function predictor	Standardized discriminant coefficient		Discriminant loading	
	Function 1	Function 2	Function 1	Function 2
Study 1[a]				
Marlowe–Crowne Scale	.90	−.96	.99	−.11
Impression Management Scale	.14	1.31	.73	.69
Study 2[b]				
Marlowe–Crowne Scale	.99	−.82	.99	−.01
Impression Management Scale	.01	1.28	.64	.77

	Standardized discriminant coefficient Function 1	Discriminant loading Function 1
Study 3[c]		
Marlowe–Crowne Scale	.71	.97
Impression Management Scale	.35	.88

Notes:
[a] $\lambda = .29$, $\chi^2(4, N = 293) = 354.96$, $p < .001$, cross-validated correct classification rate of 76.5%.
[b] $\lambda = .33$, $\chi^2(4, N = 300) = 325.93$, $p < .001$, cross-validated correct classification rate of 74.4%.
[c] $\lambda = .64$, $\chi^2(2, N = 116) = 50.18$, $p < .001$, cross-validated correct classification rate of 75.9%.

Table 3
Group descriptive statistics and effect sizes: individuals faking good and faking bad versus standard instructions.

Scale	Mean (*SD*, coefficient α)			Cohen's *d*	
	Fake good $n = 98$	Fake bad $n = 97$	Standard $n = 98$	Fake good vs standard	Fake bad vs standard
Study 1					
MCSDS	25.37 (5.43, .85)	5.86 (5.38, .86)	14.53 (4.93, .76)	2.09	1.68
IM	12.57 (4.17, .81)	2.78 (3.10, .79)	6.43 (3.49, .72)	1.60	1.11
	Fake good $n = 101$	Fake bad $n = 100$	Standard $n = 99$	Fake good vs standard	Fake bad vs standard
Study 2					
MCSDS	24.81 (6.46, .89)	5.54 (5.13, .86)	14.36 (5.22, .79)	1.78	1.72
IM	10.98 (5.10, .87)	2.52 (2.67, .74)	5.85 (3.54, .74)	1.17	1.06
	Fake good $n = 56$	Fake bad (not applicable)	Standard $n = 60$	Fake good vs standard	Fake bad vs standard
Study 3					
MCSDS	24.40 (6.28, .89)	–	15.93 (5.45, .79)	1.44	Not applicable
IM	13.04 (4.77, .87)	–	7.62 (3.43, .71)	1.30	Not applicable

1992), as they ranged from 1.11 to 2.09. In every 1 of 5 directly comparable instances, the effect size for the MCSDS exceeded the corresponding IM scale effect size. For distinguishing fake good from standard responding, the inverse-variance weighted average (Lipsey & Wilson, 2001) Cohen's *d* effect size was 1.80, $SE = 0.11$, 95% CI [1.59, 2.01], for the MCSDS and 1.35, $SE = 0.10$, 95% CI [1.16, 1.54], for the IM scale. In differentiating fake bad from standard responding, the inverse-variance weighted average Cohen's *d* effect size was 1.70, $SE = 0.12$, 95% CI [1.47, 1.93], for the MCSDS and 1.09, $SE = 0.11$, 95% CI [0.88, 1.30], for the IM scale. As such, the MCSDS and IM scale 95% confidence intervals were non-overlapping for average effect size in differentiating standard

Table 4
Adjusted effect sizes for differentiating standard from faked responding (partial η^2 for one scale statistically controlling for the other scale).

Scale	Partial η^2	
	Fake good vs standard	Fake bad vs standard
Study 1		
MCSDS	.22	.24
IM	.01	.00
Study 2		
MCSDS	.25	.26
IM	.00	.00
Study 3		
MCSDS	.08	Not applicable
IM	.01	Not applicable

responding from either faking good or faking bad and, in both differentiations, the results favored the MCSDS.

To further demonstrate the distinction between the MCSDS and IM scale in distinguishing standard and faked responding, effect sizes in Table 3 were recalculated for each of the MCSDS and IM scale while covarying out the other scale. These adjusted effect sizes are reported in Table 4 as partial eta squared values. Whereas the incremental value for adding the use of the MCSDS to the IM scale constituted a large effect size (average partial $\eta^2 = .21$), there was essentially no incremental value in adding the IM scale to the MCSDS (average partial $\eta^2 < .01$).

3.4. Receiver operating curve analyses and cut-scores

To explore MCSDS and IM cut-scores for correctly detecting fakers, we aggregated participants from our three studies to ensure groupings large enough for receiver operating characteristic (ROC) curve analyses: one group of standard responders ($n = 257$), one group faking good ($n = 255$), and one group faking bad ($n = 198$).[4] ROC analyses then evaluated separately the MCSDS and the IM scale for detecting respondents faking good versus standard responders and for detecting respondents faking bad versus standard responders. ROC curve analyses do not require assumptions of normality or homoscedasticity and are independent of prevalence rates. Diagnostic accuracy for ROC curve analysis is summarized as the area under the curve (AUC) with AUCs of .50–.70, .70–.90, or over .90 interpreted as low, moderate, or high accuracy, respectively (Fischer, Bachmann, & Jaeschke, 2003; Streiner & Cairney, 2007). AUCs can be tested for statistical significance and AUCs for two tests involving the same cases can be statistically compared (Hanley & McNeil, 1983). For diagnosing faking good versus standard responding, the AUCs (see Fig. 1) for the MCSDS and the IM scale were .891 and .815, respectively, with the value for the MCSDS being significantly larger ($z = 4.92$, $p < .01$). Cut-scores for indicating faking good that optimized the balance between sensitivity and specificity were scale scores greater than 21 for the MCSDS (sensitivity = .75; specificity = .91; relative risk ratio = 4.04) and greater than 10 for the IM scale (sensitivity = .62; specificity = .86; relative risk ratio = 1.65). For detecting faking bad versus standard responding, the AUCs (see Fig. 2) for the MCSDS and IM scale were .892 and .816, respectively, with the value for the MCSDS again being significantly larger ($z = 4.65$, $p < .01$). Cut-scores for detecting faking bad that optimized the balance between sensitivity and specificity were scale scores less than 9 for the MCSDS (sensitivity = .85; specificity = .87; relative risk ratio = 5.90) and less than 4 for the IM scale (sensitivity = .71; specificity = .79; relative risk ratio = 3.25).

[4] Although IM items were answered on 7-point Likert scales in Studies 1 and 2, and 5-point Likert scales in Study 3, standard scoring (Paulhus, 1998) of each IM item is dichotomous as either a 0 or 1, regardless of response format.

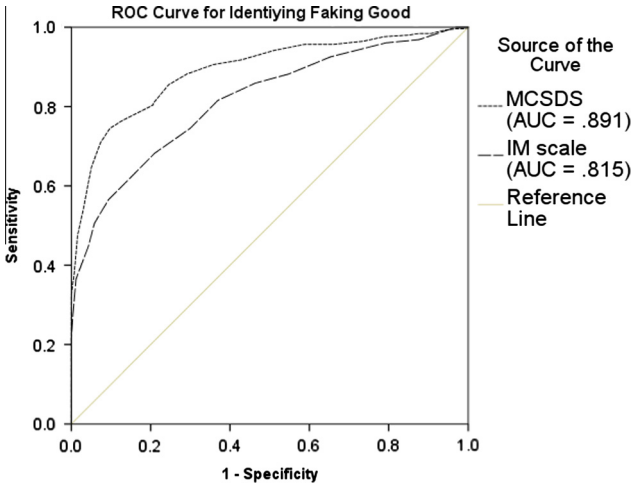

Fig. 1. Receiver operating curves for identifying fakers (faking good).

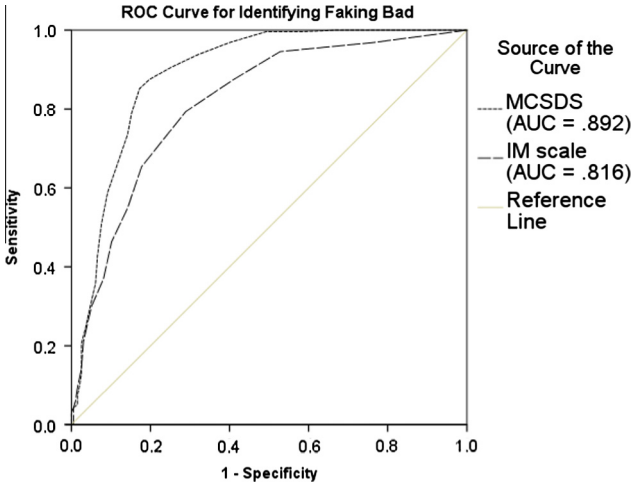

Fig. 2. Receiver operating curves for identifying fakers (faking bad).

4. Discussion

Mean values (Table 2) for the MCSDS and IM scale under standard instructions are similar to those previously reported for undergraduate students (e.g., Holden, 2007; Holden & Fekken, 1989). As such, our samples' data do not appear to be anomalous. Further, the effect sizes and identification rates for identifying fakers are similar to what have been previously reported for the MCSDS (e.g., Edens, Buffington, Tomicic, & Riley, 2001) and IM scale (e.g., Holden, 2008). Thus, our obtained differentiation between standard and faked responding also is typical.

Both the MCSDS and the IM scale have merit for the identification of fakers. Classification hit rates for identifying positive and negative dissimulators are substantially above chance although the presence of incorrect classification must certainly be recognized. That is, despite the scales' virtue for detecting fakers, non-trivial rates of misses (fakers not identified as faking) and false alarms (non-fakers flagged as fakers) do occur. Surprisingly, however, and in complete opposition to our hypothesis, analyses across each of our studies clearly indicate that the MCSDS is notably better than the IM scale at distinguishing individuals faking good and those faking bad from individuals following standard instructions. Comparisons of correct classification rates, canonical weights, canonical loadings, effect sizes, and ROC curve analyses all favored the MCSDS over the IM scale for the identification of fakers. In particular, the comparisons of AUCs demonstrated statistically

significant superiority for the MCSDS over the IM scale in detecting respondents who were faking bad and in detecting respondents who were faking good. This result is counterintuitive because the IM scale and its parent BIDR are seen as the new gold standard for measuring socially desirable responding. Further, the MCSDS is commonly considered to be outdated due to its age and its true/-false response format. The MCSDS is now more than 50 years old, and since its development "its item selection procedures have changed over time and are less applicable today than originally" (Helmes, Holden, & Ziegler, 2015, p. 25).

Can the better performance of the MCSDS relative to the IM scale be explained as a by-product of differential reliability? Consider that scores on the 33-item, true/false MCSDS tended to have greater reliability (as indexed by coefficient alpha) than scale scores based on the 20 Likert-rated, but dichotomously scored, items of the IM scale. In response to this consideration, we offer two points. First, if observed effect sizes (i.e., Cohen's *d*s) were adjusted for unreliability (e.g., divided by the square root of the estimated reliability; Grissom & Kim, 2012, p. 120), the increase in effect size would be 14% for the MCSDS (based on an average coefficient alpha under standard responding of .78) and 18% for the IM scale (based on an average coefficient alpha under standard responding of .72). If we were to calculate the differential increase in reliability for the two scales, we would thus divide 1.18 for the IM scale by 1.14 for the MCSDS scale, which yields 1.04, or approximately a 4% increase in reliability. This increase cannot account for the observed difference in unadjusted effect sizes (Table 3). As such, differences in performance between the MCSDS and IM scale cannot be primarily explained as attributable to differential reliability. Second, although corrections for unreliability can have theoretical merit, the aim of this research was to compare the MCSDS and IM scale as they are operationally used in practice for identifying distorted responding. In that regard, one plays the hand that one is dealt regardless of other factors (e.g., reliability, test length).

Although in many high-stakes assessment contexts, such as job application settings, concerns about faking center on faking good, there are a non-trivial number of situations, such as military draft and working compensation circumstances, where faking bad can be a focus. The latter have been explicitly recognized (Paulhus, 1998, p. 10) both in the development of the IM scale where scale score interpretive guidelines offer suggestions regarding both faking good and faking bad, and in the construction of the MCSDS (Crowne & Marlowe, 1960) where there was an aim to identify respondents who "present themselves in a socially desirable (or undesirable) light" (p. 350). As such, in comparing the MCSDS and IM scale as possible gold standards for identifying faking, we felt it appropriate to include, in some instances, a manipulation of faking bad.

Nevertheless, although concerns about faking bad can exist, it would seem to be of virtually no concern in many contexts, like employment or job hiring settings. Thus, in Study 3, we chose not to include a faking bad manipulation in order to demonstrate the generalizability of Study 1 and 2's results by providing a scenario that was different from Studies 1 and 2 in that it would be more common (i.e., government employment vs military draft) and would focus on a context where faking bad would not be a concern to the assessing agent; an employer would generally not worry about a job applicant who presented in order to not be hired. Further, in Study 3, we chose to use the standard 5-point rather than the also standard 7-point IM scale response format and to not offer an incentive for successful fakers. This, too, was done to demonstrate the generalizability and robustness of our finding that the new gold standard, the IM scale, does not outperform the old gold standard, the MCSDS. Although for identifying faking good in Study 3, the MCSDS and IM scale had identical classification hit rates (Table 1) and more than large effect sizes that were similar in magnitude (Table 3), the primary function standardized discriminant coefficients (Table 2) and the partial η^2 (Table 4) each strongly favored the MCSDS over the IM scale and reinforced our perspective that the newer scale (the IM scale) does not outperform the old guard (the MCSDS).

Why does the older MCSDS outperform the newer IM scale? A speculative answer is that, relative to the 20-item IM scale, the 33-item MCSDS has greater content validity associated with sampling more widely (33 items) from the domain of "behaviors which are culturally sanctioned and approved but which are improbable of occurrence" (Crowne & Marlowe, 1960, p. 350) and which, inadvertently perhaps, overlaps with the domain of faking. The number of items on a scale, as well as being able to enhance internal consistency reliability, can also broaden domain representativeness. This, subsequently, leads to the question as to why the total 40-item BIDR does not outperform either the shorter MCSDS or IM scale? A consideration here is that, in constructing the SDE and IM scales of the BIDR, there was an emphasis on the "statistical separation of the 40 items into two orthogonal item clusters" (Paulhus, 1998, p. 23). Thus, the SDE items do not extend the content validity of the IM items but, rather, sample from another domain. These speculations do require further research to evaluate their merit. For example, research that disentangles content bandwidth from scale length in the MCSDS would highlight a potentially important quality versus quantity distinction.

The current study has potential limitations. First, in our instructions, we warned about the presence of validity checks that respondents were to avoid activating. Although this information could be seen as compromising the ecological validity of our results, we gave this warning because we believe that test respondents are sophisticated and that many suspect that validity checks exist on a test regardless of whether they actually do. This conjecture is based on the growing presence of resources that coach and inform respondents on "how to beat" personality tests (e.g., Stevens-Huffman, 2012). Countering potential concerns about ecological validity, however, is that our obtained mean scale scores and effect sizes for faking are consistent with previous research (Holden, 2007, 2008; Holden & Fekken, 1989) where such warnings were not provided. Other possible limitations in our research relate to our focus on instructed faking rather than naturally-occurring dissimulation, our use of military induction and personnel selection scenarios, and our participants being undergraduate students. Comparative evaluation of the MCSDS and IM scale in contexts where natural faking occurs, such as worker compensation evaluations, job applications, and clinical settings, will serve to delimit the generalizability of the current results.

5. Conclusions

In addition to identifying cut-scores for the MCSDS and IM scale in the detection of positive and negative faking, the research presented here leads us to three important conclusions. First, consistent with previous research, both the Marlowe–Crowne Social Desirability Scale and the BIDR Impression Management Scale have substantial ability to detect respondents who are faking in their self-report. Second, the Marlowe–Crowne scale detects fakers better than the Impression Management Scale. Third, although the Marlowe–Crowne scale is more than 50 years of age, dismissal of this scale as the industry gold standard is premature.

Acknowledgment

We thank Aimy Racine and Cara Chen for their assistance with data collection.

References

Burns, G. N., & Christiansen, N. D. (2006). Sensitive or senseless: On the use of social desirability measures in selection and assessment. In M. H. Peterson & R. L. Griffith (Eds.), *A closer examination of applicant faking behavior* (pp. 113–150). Greenwich, CT: Information Age.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2: Manual for administration, scoring, and interpretation* (revised edition). Minnesota: University of Minnesota Press.

Butcher, J. N., Morfitt, R. C., Rouse, S. V., & Holden, R. R. (1997). Reducing MMPI-2 defensiveness: The effect of specialized instructions on retest validity in a job applicant sample. *Journal of Personality Assessment, 68*, 385–401.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

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

Davis, C. G., Thake, J., & Weekes, J. R. (2012). Impression managers: Nice guys or serious criminals? *Journal of Research in Personality, 46*, 26–31.

Edens, J. F., Buffington, J. K., Tomicic, T. L., & Riley, B. D. (2001). Effects of positive impression management on the Psychopathic Personality Inventory. *Law and Human Behavior, 25*, 235–256.

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. *Intensive Care Medicine, 29*, 1043–1051.

Griffith, R. L., & Converse, P. D. (2012). The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 34–52). New York: Oxford University Press.

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology, 1*, 308–311.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Routledge.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiving operating characteristic curves derived from the same cases. *Radiology, 148*, 839–843.

Helmes, E., & Holden, R. R. (2003). The construct of social desirability: One or two dimensions? *Personality and Individual Differences, 34*, 1015–1023.

Helmes, E., Holden, R. R., & Ziegler, M. (2015). Response bias, malingering, and impression management. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 16–43). New York: Academic Press.

Holden, R. R. (2000, June). *Development and preliminary data for the Holden Applicant Reliability Measure (HARM)*. Presentation at the Canadian Psychological Association annual convention, Ottawa.

Holden, R. R. (2007). Socially desirable responding does moderate scale validity both in experimental and in non-experimental contexts. *Canadian Journal of Behavioural Science, 39*, 184–201.

Holden, R. R. (2008). Underestimating the effects of faking on the validity of self-report personality scales. *Personality and Individual Differences, 44*, 311–321.

Holden, R. R. (2011). *The holden applicant reliability measure (HARM)*. Odessa, ON: Limestone Technologies.

Holden, R. R., & Book, A. S. (2012). Faking does distort self-report personality assessment. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 71–84). New York: Oxford University Press.

Holden, R. R., Starzyk, K. B., Edwards, M. J., Book, A. S., & Wasylkiw, L. (2003, June). *The holden applicant reliability measure (HARM): Construct validity in a university sample*. Presented at the Canadian Psychological Association annual convention, Hamilton, Ontario.

Holden, R. R., & Fekken, G. C. (1989). Three common social desirability scales: Friends, acquaintances, or strangers? *Journal of Research in Personality, 23*, 180–191.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. III, (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*, 348–355.

Paulhus, D. L. (1998). *Paulhus Deception Scales (PDS) user's manual*. North Tonawanda, NY: Multi-Health Systems.

Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences, 37*, 1137–1151.

Rosse, J. G., Stetcher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644.

Stevens-Huffman, L. (2012, June 11). *4 ways to beat a personality test*. <http://insights.dice.com/2012/11/06/4-ways-beat-personality-test/>.

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *Canadian Journal of Psychiatry, 52*, 121–128.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210.

Vu, A., Tran, N., Pham, K., & Ahmed, S. (2011). Reliability of the Marlowe–Crowne social desirability scale in Ethiopia, Kenya, Mozambique, and Uganda. *BMC Medical Research Methodology, 11*, 162.