PSIHOLOGIJA, 2016, Vol. 49(3), 263–276
© 2016 by the Serbian Psychological Association

UDC 159.9.072.59
DOI: 10.2298/PSI1603263S

Psychometric evaluation and short form development of the Balanced Inventory of Desirable Responding (BIDR-6)

Siniša Subotić[1,2], Strahinja Dimitrijević[3], and Sanja Radetić Lovrić[3]

[1]*Department of Psychology, Faculty of Philosophy,*
University for Business Engineering and Management,
Banja Luka, Bosnia and Herzegovina
[2]*CEON/CEES, Belgrade, Serbia*
[3]*Department of Psychology, Faculty of Philosophy,*
University of Banja Luka, Bosnia and Herzegovina

The goals of this research were to evaluate the Bosnian-Croatian-Serbian (BCS) translation of the BIDR-6 scale, develop its short form, and to present its initial convergent/discriminative validation. The sample included 827 participants. MIRT CFA analysis revealed that four-factor model (containing 32/40 items) fits the data best, with Self-Deceptive Enhancement (SDE) and Impression Management (IM) both splitting into the denial (SD-D and IM-D) and enhancement (SD-E and IM-E) factors. Fit and item properties were generally mediocre. SD-D and IM-E subscales were the strongest sources of misfit, thus SD-E and IM-D subscales were retained in the short form, which had good fit and replicated almost all main patterns of associations with other variables of interest (e.g., HEXACO personality traits) typically reported for the full SDE and IM scales in other research. Thus, 17-item BIDR-6 short form, containing only SD-E and IM-D subscales, is recommended for use in the BCS speaking area.

Keywords: Socially Desirable Responding (SDR), The Balanced Inventory of Desirable Responding (BIDR-6), Multidimensional Item Response Theory (MIRT), Confirmatory Factor analysis (CFA), HEXACO Model of Personality

Socially desirable responding (SDR) in research represents a “tendency to give answers that make the respondent look good” (Paulhus, 1991, p. 17). There is a disagreement in the literature about a potential impact of SDR. Some (Li & Bagger, 2006) have shown that SDR does not create spurious effects on the association between personality and performance variables, and that removing SDR variance does not substantially influence the criterion validity of personality measures. Others (Paunonen & LeBel, 2012) have confirmed that SDR indeed has a small impact on the criterion validity in general, but that under extreme conditions the effect can be dramatic, possibly leading to

Corresponding author: sinisasub@gmail.com

strong misrepresentation of participants' trait levels. In low stakes testing SDR is viewed as a "general method variance that is not necessarily faking and that is not necessarily substance" (Holden & Passey, 2010, p. 449). In other cases, e.g., self-reported alcohol consumption and harms research, SDR has been viewed as a significant threat to the validity (Davis, Thake, & Vilhena, 2010).

Researchers usually attempt to measure SDR using specialized questionnaires. Some questionnaires were operationalized as unidimensional measures, some as multidimensional, and some lacked a clear dimensionality (Paulhus, 1991; Uziel, 2010). The Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960) and Eysenck's L Scale (Eysenck & Eysenck, 1964) are the famous examples of unidimensional SDR measures. In recent decades, the two-dimensional SDR model was the most widely accepted, with the Balanced Inventory of Desirable Responding – BIDR/BIDR-6 (Paulhus, 1991, 1994; Paulhus & Reid, 1991) as its best known questionnaire example. The first factor of this model is associated with MMPI Alpha (general anxiety), and the second with MMPI Gamma factor (agreeableness and traditionalism) (Paulhus, 1984, 1991, 2002). Paulhus (1984, 1991) argued that these SDR factors represent self-deceptive positivity (a tendency to give an honest, but overly positive self-presentation) and impression management (self-presentation directed at others). Later, Paulhus (2002) refined this by stating that self-deception has two facets: self-deceptive enhancement (e.g.: "My first impressions always turn out to be right.", p. 58) and self-deceptive denial (e.g.: "I have never thought about killing someone.", p. 58). Paulhus (2002) suggested that trait self-enhancement, similarly to narcissism, represents 'positive illusions', which may or may not be adaptive.

Proposed techniques for SDR management (Bäckström & Björklund, 2013; Nederhof, 1985; Paulhus, 1991; Paulhus & Vazire, 2007) are fairly limited. Statistical control is arguably the most known practice and it requires the administration of some SDR scale with other measures of interest. Subjects high on SDR scale are either deleted, or their 'contaminated' scores are adjusted (Nederhof, 1985). The latter is typically done by statistical partialling of the SDR scale scores from the measures of interest in an attempt to 'purify' them (Nederhof, 1985; Paulhus & Vazire, 2007). While some authors explicitly advocate for this approach (van de Mortel, 2008), others are openly against it, suggesting that this removes valid variance instead of fixing the problem (Paulhus & Vazire, 2007, Uziel, 2010). However, even with the latter position probably being true, SDR scales are still the most convenient way for issue detection, regardless if they can be used as a 'cure'.

Recently, some authors rethought the whole concept of SDR, especially the impression management, putting forward an argument that it should be redefined as a measure of interpersonally oriented self-control, which characterizes individuals who demonstrate high self-control, particularly in social contexts (Uziel, 2010). De Vries, Zettler, and Hilbig (2014) confirmed this interpretation, showing that impression management might be an expression of Honesty-Humility HEXACO personality trait (Ashton, Lee, & de Vries, 2014; Ashton et al., 2004; Lee & Ashton, 2004). Thus, SDR questionnaires could be useful tools for the study of individual differences in general, regardless of their potential merit as an actual SDR remedy.

Research problem

BIDR-6 (Paulhus, 1991, 1994; Paulhus & Reid, 1991) is one of the most famous SDR questionnaires (Li & Bagger, 2006), typically thought of as having two dimensions (Asgeirsdottir, Vésteinsdóttir, & Thorsdottir, 2016; Bobbio & Manganelli, 2011; Hart, Ritchie, Hepper, & Gebauer, 2015; Li & Bagger, 2006; Paulhus, 1984): the Self-Deceptive Enhancement (SDE) and the Impression Management (IM). However, three factors have also been proposed (Kroner & Weekes, 1996; Paulhus & Reid, 1991), with Paulhus and Reid (1991) observing the self-deceptive content of BIDR-6 splitting into the enhancement (the claiming of positive attributes) and denial (the repudiation of negative attributes) facets. Li and Li (2008) have observed this happening to the IM scale as well, thus obtaining four factors. Paulhus and Trapnell (2009) have also argued for a four-factor model of SDR in some recent advancements. Li and Li (2008) made a remark that the BIDR-6 latent structure might be culturally dependent, requiring separate tests for different cultural backgrounds. Thus, the first goal of this research is to present a thorough psychometric evaluation of the official Bosnian-Croatian-Serbian (BCS) BIDR-6 translation, focusing on its dimensionality. Given the ambiguity in the number of factors, a confirmatory approach will be used in order to compare the plausible factor solutions. The primary framework for the analysis will be an Item Response Theory (IRT), which allows us to study how underlying latent traits interact with item characteristics, such as difficulty and discrimination (Chalmers, 2012). This useful approach has been only sporadically used with the BIDR-6 (e.g., Asgeirsdottir et al., 2016; Cervellione, Lee, & Bonanno, 2008). Most recently, Asgeirsdottir and colleagues (2016) used IRT to shorten the BIDR-6, by retaining only the best 24 items. Other authors have also developed variations of BIDR-6 short forms (e.g., Bobbio & Manganelli, 2011; Hart et al., 2015). Short forms have the advantage over a full questionnaire due to an obvious fact that many researchers might be reluctant to use a 40-item SDR measure (Hart et al., 2015). Thus, the second goal of this study is to develop the BIDR-6 short form for the BCS speaking area. Unlike Asgeirsdottir and colleagues (2016), who used a combination of the confirmatory factor analysis (CFA) and unidimensional IRT to shorten the BIDR-6, we opted to rely upon a more sophisticated multidimensional variation of IRT – MIRT, which also allows for a usage analogous to the CFA, including factor loadings and model fit estimation (Chalmers, 2012).

The third goal of this article is to present an initial insight into the convergent and discriminative validity of the BIDR-6 BCS translation. Note that our view of the SDR is more in line with the interpersonally oriented self-control perspective (de Vries et al., 2014; Uziel, 2010), than with the view of SDR measures as a way of 'weeding out bad variance'. Thus, we will primarily rely upon the findings of de Vries and colleagues (2014) as a benchmark for the BIDR-6 validation. Specifically, we expect that SDE will correlate with (low) Emotionality, Extraversion, and Conscientiousness, and that IM will correlate with Honesty-Humility, Conscientiousness, and Agreeableness, with the Honesty-Humility correlation being the strongest (de Vries et al., 2014). We also expect BIDR-6 dimensions to correlate with other measures of SDR, namely with the Brief Social Desirability Scale (Haghighat, 2007a, 2007b). Finally, it is also important to test for the gender and age differences, in order to have appropriate benchmark values for different subpopulations. In alignment with previous research, we expect that women will be higher on IM and men on SDE (e.g., Bobbio & Manganelli, 2011; de Vries et al., 2014), and that BIDR-6 dimensions will not correlate with age (de Vries et al., 2014).

Method

Sample and procedure

We used a general/convenience sample, comprised of 827 participants (58.4% women). The average age was 25.22 (*SD*=7.87) years. University students comprised 47% of the sample. Participants were recruited using online (56.6%; LimeSurvey Project Team/Carsten Schmitz, 2012) and paper-and-pen (43.4%) surveys[1]. All participants were BCS speakers (mainly from Bosnia and Herzegovina – B&H).

Measures

Balanced Inventory of Desirable Responding – BIDR-6, Form 40A (Paulhus, 1991, 1994, 2008; Paulhus & Reid, 1991). It consists of 40 Likert-type items answered on a 7-point scale (1="not true" through 7="very true"). There are 20 items per the SDE and the IM scales. Each scale also has the Enhancement and Denial subscales (10 items each). Two scoring methods exist (Paulhus, 2008; Stober, Dette, & Musch, 2002): continuous (all answers are counted) and dichotomous (only extreme answers are counted). Following recommendations from Stober and colleagues (2002) continuous scoring was used, unless noted otherwise. Adaptation to BCS included two independent back-translations. Following suggestions from the translators and student contributors, several items were slightly modified due to cultural reasons. For example, item 30: "I always declare everything at customs." was modified into: "I would always report everything at customs.", as a few people from B&H have an extensive personal experience with the customs declarations. All item translations and adaptations were verified and approved by the questionnaire's original author (D. L. Paulhus).

Brief Social Desirability Scale, Version 2 – BSDS-V2 (Haghighat, 2007a, 2007b). It contains four true-false items (two are reverse scored), which measure a single dimension of SDR. This short measure was included for a convergent validation purpose. The items were added up to create a summary score, with a 0–4 range (*Md*=*Mo*=2, *M*=1.55, *SD*=1.15). While internal consistency reliability of BSDS V2 was relatively low (*KR-20*=.64), it is comparable to typical values of other short SDR questionnaires and its own referenced value (Haghighat, 2007b).

[1] Note that the differences in SDR scores between the different data gathering modalities are out of scope of this article.

HEXACO-PI-R-60 (Ashton & Lee, 2009). This is a personality questionnaire which contains 60 Likert-type items (1="strongly disagree" through 5="strongly agree") that measure six personality traits proposed by the HEXACO psycholexical model (Ashton et al., 2014; Ashton et al., 2004; Lee & Ashton, 2004). The dimensions are: 1) Honesty-Humility (M=3.56, SD=0.65), 2) Emotionality (M=3.08, SD=0.67), 3) eXtraversion (M=3.34, SD=0.58), 4) Agreeableness (M=2.98, SD=0.59), 5) Conscientiousness (M=3.42, SD=0.59), and 6) Openness to Experience (M=3.40, SD=0.67). Internal consistency reliabilities (ω_H; McDonald, 1999) of all dimensions were above .70 (i.e., .76, .80, .75, .71, .77, and .73, respectively).

Results and discussion

We tested the dimensionality of the full BIDR-6, and performed an initial item and (sub)scale analysis, in order to establish a baseline for the short form creation and convergent/discriminative analyses. Using the Multidimensional Item Response Theory (MIRT) based Confirmatory Factor Analysis (CFA) (also known as Confirmatory Item Analysis – CIA; Chalmers, 2012), conducted in "mirt" R package (Chalmers, 2012), we compared these five models: 1) One-factor model (model conceived logically). 2) Self-Deceptive Enhancement (SDE) and Impression Management (IM) as separate factors (Asgeirsdottir et al., 2016; Bobbio & Manganelli, 2011; Hart et al., 2015; Li & Bagger, 2006; Paulhus, 1984). 3) Self-Deception-Enhancement (SD-E), Self-Deception-Denial (SD-D), Impression Management-Enhancement (IM-E), and Impression Management-Denial (IM-D) as separate factors (Li & Li, 2008; Paulhus & Trapnell, 2009). 4) IM as a single factor, but two SDE factors: SD-E and SD-D (Kroner & Weekes, 1996; Paulhus & Reid, 1991). 5) SDE as a single factor, but two IM factors: IM-E and IM-D (model conceived logically). We used two-parametric (2PL) Graded Response Model (GRM), with quasi-Monte Carlo expectation-maximization (QMCEM) algorithm for parameter calculations (Chalmers, 2012).

BIDR-6 dimensionality assessment and item properties

Initial model fits of the five tested models are shown in Table 1. Four-factor model (Model 3), clearly had the best fit. However, only RMSEA value was good, SRMSR acceptable, while the other indices were below the conventional cutoffs (Hooper, Coughlan, & Mullen, 2008). Since CFI and TLI have a tendency to penalize models with a large number of indicators per latent variable (here: 10 per factor), especially when factor loadings (Λ) are in a lower range, this discrepancy in fit values is somewhat understandable (Kenny & McCoach, 2003; Sharma, Mukherjee, Kumar, & Dillon, 2005), but the values are low nevertheless. Regardless, as the best fitting model, four-factor solution was used as a basis for further analyses.[2]

[2] The reviewers suggested a consideration of the exploratory approach. Parallel analysis (Subotić, 2013) also suggested four factors, with the exploratory analysis resulting in the same final outcome as confirmatory, thus we opted for the latter, as conceptually more appropriate in our opinion.

Table 1
The initial MIRT CFA model fits

Models	$M_2^*(df)$	CFI	TLI	RMSEA [90% CI]	SRMSR
M1: One factor	2373.14(540)	.749	.731	.064 [.061, .067]	.092
M2: SDE + IM	1958.88(539)	.806	.791	.056 [.054, .059]	.096
M3: SD-E + SD-D + IM-E + IM-D	1439.26(534)	.876	.866	.045 [.042, .048]	.070
M4: SD-E + SD-D + IM	1602.34(537)	.854	.842	.049 [.046, .051]	.071
M5: SDE + IM-E + IM-D	1923.50(537)	.810	.795	.056 [.053, .059]	.092

Note. All M_2^* tests (Cai & Hansen, 2013) were significant (*ps*<.001). As a rule of a thumb, CFI and TLI (NNFI) values ≥.95 are good, ≥.90 are acceptable; RMSEA≤.07 is acceptable, RMSEA≤.06 is good; SRMSR (SRMR) values ≤.05 are good, ≤.08 are acceptable (Hooper, Coughlan, & Mullen, 2008).

Next, it was determined that there are seven item pairs which violated local independence assumption (Chen & Thissen, 1997; Yen, 1984) on either unidimensional/univariate (each factor individually) or a multidimensional/multivariate (all four factors simultaneously) level: 3–19, 24–32, 24–38, 1–17, 11–2, 11–22, 2–39, and 8–33. Items with lower Λs in each pair (i.e., items 1, 2, 3, 8, 22, and 24) were removed. Items 13 and 14 were also removed for violating the unidimensionality assumption (Λ<|.32|; Tabachnick & Fidell, 2013, p. 654). This reduced 32-item four-factor model (referred as Model 3.1 from now on) showed an obvious fit improvement over a starting model (Model 3) according to CFI and TLI, with marginal worsening of the RMSEA and SRMS: M_2^* (298)=846.85, *p*<.001; CFI=.925.; TLI=.915.; RMSEA=.047, 90% CI [.043, .051]; SRMSR=.074. As a whole, this fit could be judged as acceptable, but mediocre (i.e., good according to RMSEA and only acceptable according to CFI, TLI, and SRMSR). Factor loadings, item thresholds (difficulties), and discriminations are given in Table 2, while factor correlations and internal consistencies according to classical test theory are given in Table 3.

Factor loadings are relatively low, with the average Λs (sums of Λ^2 are in brackets) for SD-E, SD-D, IM-E, and IM-D factors being: .59 (2.49), .52 (1.91), .51 (2.19), and .47 (2.27), respectively. Furthermore, subscales exhibit an obvious Enhancement~~Enhancement and Denial~~Denial cross-scale correlation pattern, instead of Enhancement~~Denial within-scale pattern (with especially low SD-E~~SD-D correlation), fortifying the notion that the subscales should be treated separately, and that combining them into SDE and IM scales is not advised as per our data. Internal consistencies are generally moderate (with noticeably lower value for SD-D), and roughly in line with the values expected for the BIDR-6 (Paulhus, 2008).

The majority (*n*=25) of the items has moderate unidimensional discrimination (α), several (*n*=5) have high, with items 34 and 11 having low and very high discrimination, respectively. The majority (*n*=21) of the items has low multidimensional discrimination (*MDISC*), with 11 items having moderate values. On average, both on a unidimensional and multidimensional level, SD-E has the most discriminative items (M_α=1.27, M_{MDISC}=0.75), followed by IM-E (M_α=1.05, M_{MDISC}=0.62) and SD-D (M_α=1.05, M_{MDISC}=0.62), and lastly by IM-D (M_α=0.92, M_{MDISC}=0.54).

Table 2
MIRT CFA analysis results for the Model 3.1

No.	Factor loadings (Λ)				Item thresholds						α	*MDISC*
	SD-E	SD-D	IM-E	IM-D	β_1	β_2	β_3	β_4	β_5	β_6		
$5_{\text{SD-E}}$	.64				-3.09	-2.25	-1.95	-0.95	-0.19	0.94	1.41	0.83
$7_{\text{SD-E}}$	.56				-3.41	-2.46	-1.70	-0.84	0.06	1.40	1.16	0.68
$9_{\text{SD-E}}$	.52				-2.20	-1.22	-0.22	0.69	1.68	2.84	1.04	0.61
$11_{\text{SD-E}}$	.64				-2.34	-1.46	-0.54	0.21	0.97	1.89	1.42	0.84
$15_{\text{SD-E}}$	.53				-4.19	-3.15	-1.92	-0.73	0.36	1.76	1.06	0.62
$17_{\text{SD-E}}$	.71				-3.29	-2.82	-1.64	-0.80	0.15	1.31	1.71	1.01
$19_{\text{SD-E}}$	.55				-3.84	-2.88	-1.86	-0.97	-0.22	0.85	1.11	0.65
$4_{\text{SD-D}}$		.58			-1.46	-0.43	0.30	1.05	1.94	2.83	1.20	0.71
$6_{\text{SD-D}}$		.52			-2.63	-1.43	-0.72	0.05	0.99	2.07	1.04	0.61
$10_{\text{SD-D}}$		.48			-2.37	-1.18	-0.31	0.53	1.40	2.53	0.92	0.54
$12_{\text{SD-D}}$		.52			-2.08	-1.15	-0.34	0.45	1.50	2.66	1.03	0.61
$16_{\text{SD-D}}$		.41			-2.54	-0.40	0.92	2.67	3.56	4.37	0.76	0.44
$18_{\text{SD-D}}$		.45			-1.44	0.04	0.78	1.54	2.58	3.86	0.85	0.50
$20_{\text{SD-D}}$		.67			-1.65	-0.70	-0.07	0.53	1.34	2.01	1.52	0.89
$26_{\text{IM-E}}$			.55		-2.74	-1.88	-1.08	-0.66	-0.15	0.89	1.11	0.65
$28_{\text{IM-E}}$			.49		-3.44	-2.17	-1.36	-0.55	0.21	1.46	0.95	0.56
$30_{\text{IM-E}}$			.47		-1.64	-0.81	-0.06	0.45	1.14	2.10	0.91	0.54
$32_{\text{IM-E}}$			.38		-1.58	-0.20	0.68	1.40	2.02	3.24	0.69	0.41
$34_{\text{IM-E}}$			.33		-1.53	-0.27	0.76	1.43	2.08	3.11	0.59	0.35
$36_{\text{IM-E}}$			.68		-2.27	-1.67	-1.18	-0.78	-0.38	0.40	1.57	0.92
$38_{\text{IM-E}}$			.66		-1.94	-1.46	-1.00	-0.70	-0.24	0.45	1.48	0.87
$40_{\text{IM-E}}$			.54		-2.92	-1.68	-0.79	-0.05	0.63	1.63	1.08	0.64
$21_{\text{IM-D}}$				.53	-3.00	-2.00	-1.01	-0.23	0.61	1.68	1.07	0.63
$23_{\text{IM-D}}$				.60	-1.06	-0.01	0.65	1.41	2.13	2.95	1.27	0.75
$25_{\text{IM-D}}$				.42	-0.38	0.92	1.57	2.44	3.37	4.44	0.79	0.46
$27_{\text{IM-D}}$				.50	-1.88	-0.63	0.13	0.93	1.98	3.16	0.99	0.58
$29_{\text{IM-D}}$				.46	-0.78	-0.03	0.56	1.14	1.90	2.84	0.87	0.51
$31_{\text{IM-D}}$				.48	0.09	1.02	1.46	1.93	2.61	3.46	0.92	0.54
$33_{\text{IM-D}}$				.43	-2.18	-1.34	-0.65	-0.05	0.89	2.01	0.80	0.47
$35_{\text{IM-D}}$				.47	-2.63	-1.37	-0.72	0.01	0.88	2.03	0.90	0.53
$37_{\text{IM-D}}$				.39	-2.95	-1.99	-1.34	-0.63	0.20	1.29	0.73	0.43
$39_{\text{IM-D}}$				.46	-2.27	-0.80	0.04	0.86	1.78	2.74	0.89	0.52

Note. Item thresholds represent the multidimensional item difficulty (there are *k*-1 thresholds, where *k* is the number of item ranks). α=unidimensional discrimination (calculated only for the given factor); *MDISC*=multidimensional discrimination; discrimination values below 0.34 are considered very low, 0.35–0.64 are low, 0.65–1.34 are moderate, 1.35–1.69 are high, and values over 1.70 are very high (Baker, 2001, p. 35).

Table 3
Correlations of the BIRD-6 factors (Model 3.1)

Variables	SD-E	SD-D	IM-E	IM-D
SD-E	.74	.10**	.22***	-.004
SD-D	.26***	.66	.09**	.44***
IM-E	.49***	.20***	.72	.39***
IM-D	.07	.51***	.40***	.78

Note. ** p<.01, *** p<.001. Variables were reflected so that higher values represent higher SDR. Values below the diagonal are factor correlations. Values above the diagonal are summary scores correlations. Diagonal values are the internal consistencies (ω_H; McDonald, 1999).

The majority of SD-E item thresholds are negative, with the uppermost thresholds (β_6) mostly not being too high ($M_{\beta 6}$=1.57), meaning that having a 50% probability of choosing answer 7 ("very true") requires only moderately high levels of a latent trait. Consequently, participants mostly tended to agree with the statements, i.e., SD-E items are "easy". On SD-D subscale, items 16 and 18 (which refer to the appreciation of criticism and doubting one's own abilities as a lover, respectively), were noticeably more "hard" than other SD-D items, with elevated upper thresholds (β_6 and β_5 to a degree) implying a low probability of participants strongly agreeing with the statements. Other SD-D items have slightly narrower item thresholds, grouping around the middle of a latent trait, suggesting a discrete uniform distribution of answers. IM-E also has several easier items with very low upper thresholds and narrow threshold ranges (items 36, 38, and 26), with items 30 and 40 also displaying signs of the discrete uniform distribution of answers. Finally, IM-D subscale has two obviously hard items (31 and 25, referring to stealing and revenge, respectively). Items 23, 25, 29, and 31 also fall on a harder side and items 21 and 37 on easier. Items 33, 35, and 39 display some discrete uniform distribution tendencies, but not as pronounced as the mentioned SD-D and IM-E items.

It is obvious from the analyses that both model fit and item properties of the BIDR-6 are generally mediocre, with not too many items that stand out either positively or negatively, discounting eight assumptions-violating items. Note, however, that all parameter values, while not being fully comparable due to the analyses differences, are equal to or better than the values presented in a recent IRT-based BIDR-6 analysis by Asgeirsdottir and colleagues (2016). Most obviously, item thresholds are much less extreme on our data.

Finally, if (for a comparison purpose) we estimate Model 3.1 with a conventional CFA approach (WLSMV/DWLS extraction, Theta parameterization; Beauducel & Herzberg, 2006; Rosseel, 2012) using "lavaan" program for R (Rosseel, 2012), obtained fit is noticeably worse than with the MIRT CFA: $\chi^2(458)=1752.27$, $p<.001$; CFI=.831, TLI=.817, RMSEA=.058, 90% CI [.056, .061]. This suggests that, at least on our data, MIRT approach seems to be more appropriate for the BIDR-6 analysis.

BIDR-6 short form

Using the Model 3.1 as a starting point, we proceeded with the BIDR-6 short form development, using an iterative item removal process, relying upon the Λ, *α*/MDISC, and β values, using the resulting model fit as a benchmark. After a few iterations, SD-D and IM-E subscales deteriorated, to the point that all of their items were removed. This implied that SD-D and IM-E conform with the 2PL GRM MIRT model worse than SD-E and IM-D.[3] Problems with SD-D and

[3] Individual subscale fits were not obtained as "mirt" package (Chalmers, 2012) currently cannot calculate fit when *df*s are low.

IM-E became even more apparent when we tentatively tried dichotomous scoring (Paulhus, 2008; Stober et al., 2002), after which sums of Λ^2 sharply increased for SD-E (4.33) and IM-D (6.08), but sharply decreased for SD-D (1.66) and IM-E (3.45), with Λs of four SD-D and two IM-E items dropping under |.32|. SD-D and IM-E items tended to drop out regardless of the scoring method or a number of factors used as a starting point. This happened even if the conventional CFA was used. Narrow item threshold range and a tendency for uniform discrete distribution of several SD-D and (to a lesser degree) IM-E items is probably a reason for it. Thus, we opted to remove SD-D and IM-E subscales completely, retaining only SD-E and IM-D subscales (from Model 3.1). Paulhus and Reid (1991) have shown that relative independence of enhancement and denial items is not simply a result of the item keying direction. We did retain one positive–(SD-E) and one negative-keyed subscale (IM-D), implying that keying is not a deciding factor in our case, but given the high Enhancement~~Enhancement and Denial~~Denial cross-scale correlations this might be worth exploring further via an experimental manipulation of the keying direction.

This reduced two-factor model had very good fit: M_2^* (33)=46.60, p=.059; CFI=.994.; TLI=.990.; RMSEA=.022, 90% CI [.000, .036]; SRMSR=.053.[4] Item thresholds remained very similar to the values from Model 3.1 (thus they are not shown), as did *MDISC* values for SD-E (M_{MDISC}=0.75; range: 0.60–1.20), while MDISC for IM-D slightly improved (M_{MDISC}=0.66; range: 0.54–0.95). Average Λ for SD-E did not change (M_Λ=.59; range: .51-.77; sum of Λ^2=2.48), while Λs for SD-E improved (M_Λ=.55; range: .48-.69; sum of Λ^2=3.04). Factors remained orthogonal. This model could not be improved any further, not even by a removal of two difficult IM-D items (31 and 25), whose removal (individual or joint), actually worsened the fit. Thus, we opted not to remove any more items at this point, leaving this 17-item BIDR-6 model, comprising all the SD-E and IM-D items previously included in Model 3.1, as our current recommendation for the BIDR-6 short form.

BIDR-6 convergent and discriminative validity

Given that the short BIDR-6 with only SD-E and IM-D subscales had much better fit in comparison to the best fitting full model, we tested convergent and discriminative validity using only SD-E and IM-D.

Results. Correlations of BIDR-6 factors with age, HEXACO-PI-R-60 (Ashton & Lee, 2009) personality traits, and BSDS-V2 (Haghighat, 2007a, 2007b) are given in Table 4.

[4] The MIRT CFA fit was again superior the conventional CFA fit: $\chi^2(118)=454.78$, $p<.001$; CFI=.917, TLI=.905, RMSEA=.059, 90% CI [.053, .065].

Table 4
Correlations of the BIRD-6 (short form) factors with other variables

Variables	SD-E	IM-D
Age	.18***‡	-.10**‡
BSDS	.21***	.30***
Honesty-Humility	.07*†	.47***†
Emotionality	-.31***†	.12**†
eXtraversion	.27***†	.04†
Agreeableness	-.07†	.22***†
Conscientiousness	.21***†	.22***†
Openness to Experience	.17***†	-.03†

Note. * *p*<.05, ** *p*<.01, *** *p*<.001. All variables were recoded so that higher values represent higher SDR. † marks consistent and ‡ marks inconsistent correlations with benchmark findings of de Vries and colleagues (2014). † and ‡ are not applicable for the BSDS-V2.

Men have higher SD-E scores ($M(SD)_{males}$=4.94(1.03), $M(SD)_{females}$=4.58(0.90); *t*(674.84)=-5.31, *p*<.001, *d*=-0.38), while women have higher IM-D scores ($M(SD)_{females}$=4.55(1.03), $M(SD)_{males}$=4.27(1.17); *t*(681.53)=3.60, *p*<.001, *d*=0.26). Both differences were of a small effect size (Cohen, 1992).

Discussion. SD-E and IM-D subscales alone almost perfectly replicate the correlation patterns expected for the full SDE and IM scales, making SD-D and IM-E subscale removal a non-issue. Specifically, our data replicates almost all of the hypothesized/benchmark correlations observed by de Vries and colleagues (2014) (who used full SDE and IM scales). The only obvious difference is the (very) low (Cohen, 1992) tendency of SD-E to increase, and IM-D to decrease with age in our sample, but this probably does not require separate age-group norming. Slight caveat should also be put on SD-E~~Honesty-Humility and SD-E~~eXtraversion correlations. In the first case, there is a significant positive correlation, while the benchmark value was .00. However, given that our correlation fall under a trivial effect size (Cohen, 1992), the values are still comparable. In the second case, benchmark correlation is somewhat higher (.46 versus .27). All other correlations fall in line almost perfectly with the benchmark values. This includes the IM-D~~Honesty-Humility as the strongest observed correlation, thus confirming the previous notion that impression management might be a partial expression of Honesty-Humility (de Vries et al., 2014) and/or interpersonally oriented self-control (Uziel, 2010). More elaborated investigation of the underlying reasons for this is out of scope of this article, but it is obvious that the BIDR-6 short form presented here could be used for such investigation in the BCS speaking area.

SD-E and IM-D correlate with a short measure of SDR (BSDS-V2), as it was conceptually expected, even though correlations were in a lower range (Cohen, 1992). Finally, SD-E and IM-D also replicated the typical IM/SDE gender trends (Bobbio & Manganelli, 2011; de Vries et al., 2014), with women having higher IM-D and men SD-E. However, given the small effect sizes, separate gender norms might not be needed for low stake testing.

General discussion

"Controversy over the dimensionality of SDR is ongoing" (Hart et al., 2015, p. 7). This includes the BIDR-6. According to our data, the best fitting BIDR-6 structure was a four-factor one, with SDE and IM scales each splitting into the enhancement (SD-E & IM-E) and denial (SD-D & IM-D) dimensions. Paulhus and Reid (1991) also obtained enhancement-denial split for SDE, but not for IM scale. Four factors have been theorized by Paulhus and Trapnell (2009) in recent SDR model conceptualizations, but, until now, four-factor BIDR-6 structure was observed only in a Chinese sample (Li & Li, 2008).

Even though (after the deletion of eight assumptions-violating items) psychometric properties of the items in our four-factor model were equal to or better than a recent IRT-based BIDR-6 analysis (Asgeirsdottir et al., 2016), values were still mediocre, as was the overall model fit. SD-D and IM-E subscales were the causes for less-than-good fit, probably due to narrow item threshold ranges and some tendency towards the discrete uniform distribution of answers. This made the process of shortening the BIDR-6 very easy, as simply removing SD-D and IM-E and retaining the SD-E and IM-D subscales (17 items in total) produced a good fit. Several more items could have been removed due to their difficulty, but this offered no fit improvement advantages, and we argue that no further item removal is advised until a performance of the retained items is investigated under a deliberate faking condition (Asgeirsdottir et al., 2016). Such investigation would be an obvious next research step.

We also presented an evidence of multidimensional IRT-based CFA (i.e., CIA) analysis having clear fit advantages over a conventional CFA for the BIDR-6. Thus, we advise other researchers to consider using MIRT CFA. Expected item properties would likely still be mediocre, but since SDR represents either a method variance (Holden & Passey, 2010) or a response style (of a sort) measuring the interpersonally oriented self-control (de Vries et al., 2014; Uziel, 2010), it is unrealistic to expect anything better from the BIDR-6 (or other SDR scales).

It appears that the removal of SD-D and IM-E subscales did not compromise the validity of BIDR-6, as SD-E and IM-D retain almost all of the convergent and discriminative properties expected for the full SDE and IM scales. Namely, SD-E and IM-D replicate the benchmark correlations with the HEXACO personality traits, most importantly, with the Honesty-Humility dimension, making this short form fully suitable for the further investigation of the hypothesis that IM might be an expression of Honesty-Humility (de Vries et al., 2014). The short form also replicates the typical gender patterns (Bobbio & Manganelli, 2011; de Vries et al., 2014) and correlates (albeit lower) with at least one other SDR measure (Haghighat, 2007a, 2007b). It only slightly differs from the benchmark in regards to age-related trends (de Vries et al., 2014).

In conclusion, the short form of this BIDR-6 BCS translation has sufficiently adequate psychometric properties to be used for general research purposes and is recommended over the long form.

References

Asgeirsdottir, R. L., Vésteinsdóttir, V., & Thorsdottir, F. (2016). Short form development of the Balanced Inventory of Desirable Responding: Applying confirmatory factor analysis, item response theory, and cognitive interviews to scale reduction. *Personality and Individual Differences, 96,* 212–221. doi:10.1016/j.paid.2016.02.083

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345. doi:10.1080/00223890902935878

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A review of research and theory. *Personality and Social Psychology Review, 18*(2), 139–152. doi:10.1177/1088868314523838

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., ... De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*(2), 356–366. doi:10.1037/0022–3514.86.2.356

Bäckström, M., & Björklund, F. (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology, 54*(2), 152–159. doi:10.1111/sjop.12015

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). University of Maryland College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Beauducel, A., & Herzberg, P. (2006). On the Performance of Maximum Likelihood versus Means and Variance Adjusted Weighted Least Squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(2), 186–203.

Bobbio, A., & Manganelli, A. M. (2011). Measuring social desirability responding. A short version of Paulhus' BIDR 6. *Testing, Psychometrics Methodology in Applied Psychology, 18*(2), 117–135. Retrieved from http://goo.gl/0xSsEo

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*(2), 245–276. doi:10.1111/j.2044–8317.2012.02050.x

Cervellione, K. L., Lee, Y. S., & Bonanno, G. A. (2008). Rasch modeling of the self-deception scale of the balanced inventory of desirable responding. *Educational and Psychological Measurement, 69*(3), 438–458. doi:10.1177/0013164408322020

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. Retrieved from https://goo.gl/nOqzOv

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289. doi: 10.2307/1165285

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. doi:10.1037/0033–2909.112.1.155

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4) 349–354. doi: 10.1037/h0047358

Davis, C. G., Thake, J., & Vilhena, N. (2010). Social desirability biases in self-reported alcohol consumption and harms. *Addictive Behaviors, 35*(4), 302–311. doi:10.1016/j.addbeh.2009.11.001

de Vries, R. E., Zettler, I., & Hilbig, B. E. (2014). Rethinking trait conceptions of social desirability scales: Impression Management as an expression of Honesty-Humility. *Assessment, 21*(3), 286–299. doi:10.1177/1073191113504619

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

Haghighat, R. (2007a). The development of the brief social desirability scale (BSDS). *Europe's Journal of Psychology, 3*(4). Retrieved from http://goo.gl/4bG6jV

Haghighat, R. (2007b). *Brief social desirability scale version 2 scoring manual.* Unpublished manuscript.

Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The Balanced Inventory of Desirable Responding short form (BIDR-16). *SAGE Open*, *5*(4), 1–9. doi:10.1177/2158244015621113

Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences, 49*(5), 446–450. doi:10.1016/j.paid.2010.04.015

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53–60. Retrieved from http://goo.gl/NfO8SD

Kenny D. A., & McCoach D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*(3), 333–351. doi:10.1207/S15328007SEM1003_1

Kroner, D. G., & Weekes, J. R. (1996). Balanced inventory of desirable responding: Factor structure, reliability, and validity with an offender sample. *Personality and Individual Differences, 21*(3), 323–333. doi:10.1016/0191–8869(96)00079–7

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 39*(2), 329–358. doi:10.1207/s15327906mbr3902_8

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14*(2), 131–141. doi:10.1111/j.1468–2389.2006.00339.x

Li, F., & Li, Y. (2008). The Balanced Inventory of Desirable Responding (BIDR): A factor analysis. *Psychological Reports, 103*(3), 727–731. doi:10.2466/pr0.103.3.727–731

LimeSurvey Project Team / Carsten Schmitz. (2012). LimeSurvey: An open source survey tool [Computer Software]. Hamburg, Germany: LimeSurvey Project. Retreived from https://www.limesurvey.org/

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263–280. doi:10.1002/ejsp.2420150303

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598–609. doi:10.1037/0022–3514.46.3.598

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

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. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, & D. N. Jackson (Eds.), *Role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Lawrence Erlbaum.

Paulhus, D. L. (2008). *BIDR Version 6 – Form 40A.* Unpublished manuscript.

Paulhus, D. L., & Reid, D. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60*, 307–317. doi:10.1037/0022–3514.60.2.307

Paulhus, D. L., & Trapnell, P. D. (2009). Self-presentation of personality: An agency–communion framework. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality psychology* (pp. 493–517). New York, NY: Guilford Press.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R.W. Robins, R.C. Fraley & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York, NY: Guilford Press.

Paunonen, S. V., & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology, 103*(1), 158–175. doi:10.1037/a0028165.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. Retreived from http://goo.gl/l4kgYS

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research, 58*(7), 935–943. doi:10.1016/j.jbusres.2003.10.007

Stober, J, Dette, D. E., & Musch, J. (2002). Comparing continuous and dichotomous scoring of the Balanced Inventory of Desirable Responding. *Journal of Personality Assessment, 78*(2), 370–389. doi:10.1207/S15327752JPA7802_10

Subotić, S. (2013). Pregled metoda za utvrđivanje broja faktora i komponenti (u EFA i PCA). *Primenjena psihologija, 6*(3), 203–229.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Uziel, L. (2010). Rethinking social desirability scales from impression management to interpersonally oriented self-control. *Perspectives on Psychological Science, 5*(3), 243–262. doi:10.1177/1745691610369465

van de Mortel, T. F. (2008). Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, 25*(4), 40–48. Available from http://goo.gl/CGukFt

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145. doi:10.1177/014662168400800201

RECEIVED 25.03.2016.
REVISION RECEIVED 22.08.2016.
ACCEPTED 23.08.2016.