Assessment
2016, Vol. 23(5) 588–602

DOI: 10.1177/1073191115590852
asm.sagepub.com
SAGE

Quantifying the Association of Self-Enhancement Bias With Self-Ratings of Personality and Life Satisfaction

Daniel Leising[1], Kenneth D. Locke[2], Elena Kurzius[1,3], and Johannes Zimmermann[4]

Abstract

Kwan, John, Kenny, Bond, and Robins conceptualize self-enhancement as a favorable comparison of self-judgments with judgments *of* and *by* others. Applying a modified version of Kwan et al.'s approach to behavior observation data, we show that the resulting measure of self-enhancement bias is highly reliable, predicts self-ratings of intelligence as well as does actual intelligence, interacts with item desirability in predicting responses to questionnaire items, and also predicts general life satisfaction. Consistent with previous research, however, self-ratings of intelligence did not become more valid when controlling for self-enhancement bias. We also show that common personality scales like the Rosenberg Self-Esteem Scale reflect self-enhancement at least as strongly as do scales that were designed particularly for that purpose (i.e., "social desirability scales"). The relevance of these findings in regard to the validity and utility of social desirability scales is discussed.

Keywords

social desirability, self-enhancement bias, self-report, social relations model, perceiver, target, evaluation

The Problem of Socially Desirable Responding

Most studies in the field of personality research rely exclusively on self-report measures (Vazire, 2006). For decades, one of the reasons why this approach to personality assessment has been criticized is the suspicion that people who respond to questionnaires may tend to portray themselves in overly positive ways. More important, the extent to which people portray themselves in overly positive or negative ways may *differ* (John & Robins, 1994). Terms like "self-enhancement/-derogation" or "socially (un-)desirable responding" are often used when referring to such interindividual differences. They denote the extent to which a person description implies a positive—or a negative—evaluation of a target person that is not justified by how the person "actually is," but rather lies in the eye of the beholder. Socially (un-)desirable responding may be a problem in at least two ways: First, it may weaken the criterion validity of measures. For example, selection procedures may favor targets who present themselves better over targets who actually "are" better. Second, it may weaken the discriminant validity of measures by inducing correlations that only reflect the shared susceptibility of different measures to the same bias, rather than actual associations between the targets' traits. For these reasons, it is important to make socially desirable responding itself measurable, and to determine how strongly it affects the self-report scales that are typically used in psychological research.

Aims of the Present Study

In this study, we present an approach for measuring socially desirable responding that is strongly inspired by an important conceptual contribution by Kwan, John, Kenny, Bond, and Robins (2004). Their work is presented below, along with a few important methodological modifications that we undertook in the present study. After introducing our method, we investigate the characteristics of the resulting measure of socially desirable responding in several important regards, using an empirical data set. In particular, we ask (a) whether the measure is reliable, (b) whether it actually measures bias as opposed to "actual behavior," (c) whether it shows the properties of a moderator or

[1]Technische Universität Dresden, Dresden, Germany
[2]University of Idaho, Moscow, ID, USA
[3]Martin-Luther-Universität Halle-Wittenberg, Halle, Germany
[4]Universität Kassel, Kassel, Germany

Corresponding Author:
Daniel Leising, Department of Psychology, Technische Universität Dresden, 01062 Dresden, Germany.
Email: daniel.leising@tu-dresden.de

suppressor variable impairing validity, (d) to what extent self-reports of personality and life satisfaction are "contaminated" with this bias, (e) to what extent so-called "lie scales" or "social desirability scales" reflect the bias (which they are supposed to), and (f) to what extent associations between the bias and self-reports of personality are moderated by the evaluativeness of the items that are used for assessing the latter. All of these are crucial issues that have featured in the literature on socially desirable responding for decades. To foster the reader's understanding of these issues, we will now present a brief overview of this literature, and of the major points of empirical and conceptual progress therein.

Overview of the Literature

McCrae and Costa (1983; cf. Wiggins, 1973) used the terms *substance* and *style* to distinguish the target's actual characteristics (= substance) from the positive or negative presentation of those characteristics (= style). Indisputably, people differ in how socially desirable their actual behaviors are (Hofstee & Hendriks, 1998). In contrast, *socially desirable responding* (e.g., to a questionnaire) is a style issue: It concerns the evaluative "spin" that is present in how the target's personality is presented. Most person judgments probably incorporate both substance and style.

Early attempts at identifying persons who tend to overestimate their virtues and underestimate their weaknesses used so-called "lie scales" or "social desirability scales" (e.g., Crowne & Marlowe, 1960). In these scales, respondents are presented with lists of personality features that are either desirable but rare, or undesirable but common. Self-reports of persons who systematically endorse the first but deny the second kind of personality features in their self-descriptions are identified as biased because such response patterns are considered "too good to be true." Lie scales are supposed to capture style, but not substance. If this was what they did, then using a lie scale as a covariate should improve the correlation between self-ratings and criterion variables (e.g., other-ratings), because then the targets' tendencies to judge themselves too leniently or harshly would be controlled for. However, empirical studies could not confirm this prediction (Borkenau & Ostendorf, 1992; McCrae & Costa, 1983; Piedmont, McCrae, Riemann, & Angleitner, 2000). To the contrary, self–other agreement often *dropped* considerably when lie scales were used as covariates. One of the explanations that were offered for such findings was that lie scales may partly capture differences in how desirably people actually behave (McCrae & Costa, 1983), so partialling them out would remove substance variance as well. In a recent review, McGrath, Mitchell, Kim, and Hough (2010) came to the conclusion that studies consistently failed to show increases in self-report validity when lie scales were used as covariates, and thus the use of lie scales, despite being common, is still lacking empirical support. Other authors, however, challenged such conclusions, presenting evidence that some of the scales devised to assess response styles do react in predictable ways when participants are instructed to "fake good" or "fake bad" (e.g., Baity, Siefert, Chambers, & Blais, 2007; Morey & Lanier, 1998), that such effects may in fact be accompanied by decreases in predictor–criterion correlations (J. L. Anderson, Sellbom, Wygant, & Edens, 2013), and that in some instances the partialling out of lie scales *does* improve validity (e.g., Rohling et al., 2011). The dispute has not been ultimately resolved yet, and the debate continues (e.g., McGrath, Kim, & Hough, 2011; Morey, 2012). The issue is complicated by the fact that "faking" research addresses only intentional response distortions whereas socially desirable responding may come about both intentionally and unintentionally (Paulhus, 1984; Sackeim & Gur, 1979). In the present article, we will not attempt to resolve this issue, but rather will introduce and test a novel method of *measuring* self-enhancement bias (SEB) that goes beyond mere self-report.

Paulhus (2002) emphasized that accurately assessing the style component of a person judgment is only possible if a measure of the target's "true" features (= substance) is available. The extent to which a perceiver deviates from that truth in his or her description of a target would then be interpretable as bias. If no such criterion variable is available, the relative extent to which a measure reflects substance versus style has to remain unclear. Paulhus and John (1998) used judgments by their targets' acquaintances as measures of the targets' actual personalities and interpreted the residuals from linear regressions of the targets' self-ratings on these informant-ratings in terms of bias. In the present study, we use a similar approach (see below).

Important conceptual progress was also made with a paper by Kwan et al. (2004), who pointed out that overly positive (or negative) self-descriptions may reflect either of *two* comparisons: First, the targets' self-ratings may be more (or less) positive or negative than the same targets' judgments *of* the average other person (*social comparison bias*). Second, the targets' self-judgments may be more (or less) positive than judgments of the same targets *by* the average other person (*self-insight bias*). Only considering one of these possibilities will always leave one possible confounder unaccounted for: People who judge themselves more positively than they judge others may actually "be better" than those others, thus it is necessary to also control for how others perceive them. On the other hand, people who judge themselves more positively than they are judged by others may simply expose a general favorability bias in judgments of *all* targets. Thus, in order to assess people's evaluative attitudes toward *themselves in particular*, both types of comparison have to be considered.

In their study, Kwan et al. (2004) let 128 student participants get acquainted with one another in small groups of 4 or 5 persons, and then provide self- and other-ratings in a round robin format. Analyses showed that the different kinds of positivity bias (i.e., comparisons of self-ratings with ratings of others, by others, and of and by others) were sufficiently distinct from each other, and also differentially associated with measures of "adjustment" (e.g., self-esteem). These findings confirm that separately analyzing different kinds of evaluative comparisons between self and others is in fact necessary. Kwan et al.'s (2004) model, which essentially represents an application of the social relations model (Kenny, 1994) to the issue of self-enhancement, constitutes the conceptual backbone of the present article: Our measure of self-enhancement also reflects the positivity of a person's self-judgment that remains after controlling for how positively the person judges others and is judged by others. However, we introduce a number of methodological refinements to Kwan et al.'s approach, which will be explained in the "Method" section.

Several researchers have argued that there must be a close correspondence between the social desirability of an item's content (i.e., how positive endorsing the item makes the target appear), and the perceiver's evaluative attitude toward the target (Bäckström, Björklund, & Larsson, 2009; John & Robins, 1993; Leising & Borkenau, 2011; McCrae & Costa, 1983; Saucier, 1994; Vazire, 2010). This correspondence may be conceptualized as an interaction effect: If a perceiver does not have a particularly positive or negative attitude toward a target, then it should not matter much whether the item she or he uses has a positive or a negative evaluative connotation—the perceiver may or may not endorse items, irrespective of how "good" or "bad" doing so will make the target appear. Likewise, even if the perceiver does have a strong evaluative attitude toward the target, that attitude should only translate into her or his ratings of the target if the respective item also has an evaluative connotation. If the item is evaluatively neutral, the responses of perceivers with different evaluative attitudes toward the same target should not differ much. Although theoretical deliberations such as these seem compelling, empirical research has only recently begun to directly investigate their validity (e.g., Bäckström et al., 2009; Leising, Gallrein, & Dufner, 2014). Therefore, we also test whether our new measure of SEB interacts with the evaluative tone of the items in predicting responses to self-report questionnaires.

Method

Sample

The target sample in this study comprised 100 women and 101 men with a mean age $M = 24.09$ years ($SD = 5.05$). The average level of education in the sample was high, as 174 (86.6%) participants had attained "Abitur" (the highest secondary education degree in Germany, which is attained by less than 50% of a birth cohort). Participants were recruited among the students of a midsized university in the East of Germany, and from the local community. They were paid 30 Euro for their participation.

Procedure

Our measure of self-enhancement is based on the comparison of people's self-judgments with how they judge others and how they are judged by others (Kwan et al., 2004). To enable these comparisons, participants judged their own behavior in a set of standardized laboratory situations. They also judged the behavior of four so-called "standard targets" in the same situations, and were judged by four so-called "standard perceivers."

On arriving at the lab, the participants were greeted by the experimenter and asked to fill out a consent form. Afterward, they completed a set of self-report personality questionnaires and three brief intelligence tests (see below). Next, the experimenter presented them with the following series of 17 tasks: (1) reading a brief meteorological text (explaining the average temperature curve in the course of a day); (2) describing a book or a movie they enjoyed; (3) inventing as many different uses as possible for a cork; (4) inventing a brief story based on an image displayed on a TAT (thematic apperception test) card; (5) indicating the years in which World War II began and ended, and the year in which the Berlin wall was built; (6) explaining the meaning of the word "symmetry"; (7) calculating the square of 16 and the square root of 121; (8) reporting on some recent experience of being "successful"; (9) explaining what is important in life and what they would like to achieve; (10) recounting a recent experience of being angry or sad; (11) discussing a few things that they worry about; (12) recounting a recent experience of "simply having a good time"; (13) talking about something that they always wanted to try, but did not dare to try yet; (14) telling a joke of their own choice; (15) singing a song of their own choice; (16) taking part in an assertiveness role-play (i.e., calling a neighbor late in the evening, to complain about loud music coming from her house); and (17) pantomiming the word "party" (which in German only means "festivity," not "political party"). Some of these tasks were inspired by, or directly adapted, from previous studies investigating similar issues (e.g., Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004). The tasks were always presented in the same order. The participants' behavior during the complete sequence of tasks was videotaped (i.e., 1 video-clip per participant was recorded).

After completing all tasks, the participants were asked to judge five videotapes, using a list of adjectives (see below). Each videotape showed a person engaging in the 17 tasks

described above. One of these persons was the current participant himself or herself (i.e., we showed each participant the videotape what we had just recorded with him or her), the other four persons were the four standard targets, which were the same for all participants. The participants were to judge the *behavior* of each of the five persons, using the adjectives. The five tapes were presented in random orders. The videotapes of all participants were also judged by the four standard perceivers, which were the same for all participants. Thus, each participant was part of a group of 201 perceivers who judged the same four standard targets, and also part of a group of 201 targets who were judged by the same four standard perceivers. The ratings were to reflect the respective target's behavior across the 17 situations, that is, each perceiver judged each target only once, but on the complete set of adjectives.

This design enabled us to assess (a) how positively the participants judged others (i.e., the standard targets), (b) how positively the participants were judged by others (i.e., the standard perceivers), and (c) how positively the participants judged themselves. These are the three kinds of information that are needed to compute the SEB according to Kwan et al. (2004). The overall positivity of a judgment was quantified in terms of the profile correlation between a perceiver's ratings of a target on the adjective list and average ratings of those adjectives' social desirability (see below). Note that no personal interaction of any kind took place between perceivers and targets in the present study—this study was only concerned with issues of social judgment, not social interaction.

The four *standard targets* had been assessed before we started recruiting the 201 participants. They were selected from a larger group of 20 pilot targets, all of whom engaged in the same set of tasks to which the 201 participants were later exposed. Two men and two women of about equal age (about 25) were selected as standard targets. We took care to pick standard targets that differed considerably in how they engaged with the tasks. In order to reduce the risk that the participants knew the standard targets, the latter were recruited from a town about 30 kilometers away from the city where the main study was conducted. Prior acquaintance between perceivers and targets was to be avoided, in order to rule out effects of perceiver loyalty (e.g., a "pal-serving bias" and range restriction; Leising, Erbs, & Fritz, 2010; Leising et al., 2014; Peabody & Goldberg, 1989).

The four *standard perceivers* provided their ratings at the very end of the study, after the laboratory assessments of the 201 participants had been completed. We selected two men and two women of about equal age (about 25) as standard perceivers. For their ratings of the participants' behavior, they used the same list of adjectives that the participants had used for rating themselves and the four standard targets.

Measures

Adjective List. We used a list of 46 adjectives (see the appendix) for all judgments of behavior in the lab. The same list was used for the participants' ratings of themselves, the participants' ratings of the standard targets, and the standard perceivers' ratings of the participants. The behavior ratings were needed for computing the SEB (see below). In addition, the participants also completed a more traditional self-report version of this measure before engaging with the tasks, that is, they provided general retrospective descriptions of their own personalities by means of the adjectives. The adjective list comprised the 30 terms that Borkenau and Ostendorf (1998) had compiled as a brief measure of the Big Five personality factors (each factor is assessed by three positive and three negative items), and 16 adjectives selected from the interpersonal adjective list (Jacobs & Scholl, 2005), such that each of the eight "octants" of the interpersonal circumplex model (Wiggins, 1979) was assessed by two items.

Intelligence Tests. Three brief intelligence tests were used to assess the participants' intellectual capacities. The results of these tests were needed as an accuracy criterion in the suppressor and moderator analyses that we present below. The digit linking test (Zahlen–Verbindungs test; Oswald & Roth, 1987) is a test of *cognitive processing speed* that uses numerical material. In this test, participants use a pencil to draw connections between 90 circles containing the numbers 1 to 90. The circles are to be connected in ascending order. Performance is measured in terms of the time it takes a participant to complete this task. Due to limited resources, we only used two of the four standard sheets this test comprises. Studies have shown good retest reliability (r = .84; interval: 6 months) for the test score (Oswald & Roth, 1987) and a correlation of r = .71 with general intelligence (Vernon, 1993).

Subtest 3 of the performance test system (Leistungsprüfsystem; Horn, 1983) was used to assess *Reasoning* or *Fluid Intelligence* (Cattell, 1971; Thurstone, 1938). This test comprises 40 sets of 8 geometrical figures. For each item, participants are to decide which of the figures does not conform to the logical pattern shared by the other seven figures (e.g., the first item consists of a circle and seven diamonds). The items become increasingly difficult. Performance is measured in terms of the number of correct responses. Studies have shown acceptable retest reliability (r = .66; intervals varying) for this test, and a correlation of r = .83 with general intelligence (Horn, 1983).

Finally, we used the multiple choice vocabulary intelligence test (Mehrfachwahl–Wortschatz test, MWT-B; Lehrl, 2005) to assess crystallized intelligence (Cattell, 1971). In this test, the participants are to identify real words among nonwords that only look like real words. Each item contains one real word and four distractors, and items become

Table 1. Intercorrelations of Self-Report Scales, Self-Evaluation Factors, and Measures of Judgment Positivity.

				Positivity			
Scale	α	Factor 1: PSR	Factor 2: CTL	Self-ratings	Standard perceivers	Standard targets	Self-enh bias
Factor 1: PSR	—	1.00	.28	.51	.28	−.03	.43
Factor 2: CTL	—	.28	1.00	.31	.13	−.14	.29
BDI	.85	−.90	−.15	−.41	−.24	−.02	−.34
RSE	.86	.84	.29	.43	.27	−.01	.36
LOT-R	.76	.73	.38	.45	.37	−.03	.34
NPI	.89	.37	.95	.33	.12	−.17	.32
BIDR-IM	.63	.31	−.17	.26	.08	.09	.24
BIDR-SDE	.63	.57	.33	.30	−.03	−.06	.34
BIDR-total	.69	.54	.07	.35	.04	.03	.36
SES-17	.69	.43	.01	.25	.05	.03	.25

Note. PSR = Positive Self-Regard; CTL = Claim to Leadership; BDI = Beck Depression Inventory (modified); RSE = Rosenberg Self-Esteem Scale; LOT-R = Life Orientation Test–Revised; NPI = Narcissistic Personality Inventory; BIDR = Balanced Inventory of Desirable Responding; SDE = Self-Deceptive Enhancement; IM = Impression Management; SES-17 = Soziale Erwünschtheits-Skala (Social Desirability Scale). $|r| > .14$ are significant at $p < .05$.

increasingly difficult. Again, performance is measured in terms of the number of correct solutions. Studies have shown good retest reliability ($r = .87$, interval: 14 months) for this test, and a correlation of $r = .81$ with general intelligence (Lehrl, 2005).

We computed an overall intelligence score, by averaging the standardized scores of the three individual tests. Before doing this, completion times for the digit linking test were multiplied with −1, such that higher scores also reflected higher ability. The internal consistency of the total intelligence score was $\alpha = .60$, indicating that it covered a relatively broad measurement domain.

Personality Questionnaires and Social Desirability Scales. Before engaging in the 17 lab tasks, the participants were asked to complete a number of well-established self-report questionnaires. We picked questionnaires that had a strong self-evaluative component to them, as we expected that their items would at least partly reflect socially desirable responding. In addition, we also presented the participants with two measures that are *explicitly* aimed at capturing socially desirable responding. All items were to be answered using a 5-point scale that ranged from 1 = *does not fit at all* to 5 = *fits perfectly*. The only exception were the items of the modified Beck Depression Inventory (BDI), for which the response scale ranged from 1 = *never* to 5 = *almost always*. The internal consistencies of the scales are displayed in Table 1.

Adjective list. As noted above, the same items that the participants used for judging their own and others' behavior in the lab also appeared as a traditional retrospective self-report measure of personality. Only two items pertaining to intellectual capacity were used in subsequent analyses (see below).

Beck Depression Inventory. The BDI (Beck & Steer, 1987) asks participants to report how much they were recently affected by some of the most prominent symptoms of depression. In the present study, we used a modified German version of the BDI (Schmitt & Maes, 2000) with only 20 items (excluding weight loss). In this version, the response format is changed such that participants are asked to report the frequency of each symptom, rather than deciding which of four alternative sentences describes themselves best (as is the case in the original BDI).

Rosenberg Self-Esteem Scale. The Rosenberg Self-Esteem Scale (Rosenberg, 1965) consists of 10 items assessing the overall positive or negative views that people have of themselves. In the present study, the revised German version by Collani and Herzberg (2003; Ferring & Filipp, 1996) was used.

Life Orientation Test–Revised (LOT-R). The German translation of the original version (Scheier, Carver, & Bridges, 1994) of this measure was devised by Glaesmer, Hoyer, Klotsche, and Herzberg (2008). The measure assesses dispositional optimism (i.e., expecting the best for one's personal future), and comprises six items only (plus four filler items).

Narcissistic Personality Inventory. The Narcissistic Personality Inventory (NPI) was developed in reference to the DSM-III (APA, 1980) criteria for Narcissistic Personality Disorder. The original measure (Raskin & Hall, 1979; Raskin & Terry, 1988) comprises 40 forced-choice items: Each item contains two alternative sentences and participants are to decide which of the two describes themselves better. In the present study, we used the German translation

of the NPI by Schütz, Marcus, and Sellin (2004). To make the NPI better comparable to the other measures, we presented all 80 sentences as separate items.

Balanced Inventory of Desirable Responding (BIDR). The German translation of the original version (Paulhus, 1984, 1991) of this measure was devised by Musch, Brockhaus, and Bröder (2002). Its 20 items are supposed to assess two kinds of socially desirable responding, "self-deceptive enhancement" (SDE; involuntary; actually having an overly positive view of oneself) and "impression management" (IM; voluntary; presenting oneself in an overly positive fashion, but not really believing that image of oneself to be true). The measure is essentially based on the "too good to be true" logic discussed above. In the present study, the correlation between the two subscales was $r(199) = .23$, $p = .001$.

Social Desirability Scale (SES-17). This German measure was devised by Stoeber (1999) and is also supposed to assess socially desirable responding based on the traditional "too good to be true" logic. It comprises 17 items assessing rare virtues and common flaws.

Life Satisfaction Ratings. The last 115 participants that took part in the present study were asked to report how satisfied they were with 21 different aspects of their own lives. Specifically, they were asked to rate their level of satisfaction with their own physical health, work conditions, intimate relationship (if they had one), education, outer appearance, income, overall financial situation, the German legal system, the German political system, the current German government, their friends, sexual life, living situation, body weight, style of dress, parents, siblings, children, the TV program in general, as well as their leisure activities and future prospects. The participants used a scale ranging from 1 (*not at all satisfied*) to 5 (*very satisfied*) for these ratings. We included life satisfaction ratings because we were interested in the degree to which such ratings reflect SEB.

Item Desirability Ratings

A separate group of 30 participants (22 female; age: $M = 22.03$ years, $SD = 3.7$) judged all items of the personality questionnaires, the adjective list, and the social desirability scales with regard to how much of a positive or negative evaluation is implied by using the respective description for a target person. These ratings were essential for computing the self-enhancement index, and for investigating the interaction between perceiver attitude and evaluative item tone (see below). The response options for these ratings ranged from 1 (=*very negative*) to 5 (=*very positive*). Interrater reliability was ICC(2, 30) = .99.

Computation of the Self-Enhancement Bias

Based on Kwan et al.'s model, we conceptualized the SEB as the tendency to judge oneself more positively than one judges others, *and* more positively than one is judged by others. We operationalized judgment positivity in terms of profile correlations between a perceiver's judgments of a target on the 46 adjectives and averaged ratings of these items' social desirabilities (Edwards, 1953; Leising et al., 2010; Locke & Horowitz, 1997). The main advantage of this approach is that it exactly quantifies the extent to which the perceiver's ratings of the target may be predicted from how positive or negative they make the target appear: A profile correlation of 1 implies that the perceiver rated the target exactly as one would rate a "perfect person" (i.e., the more positive/negative the item, the more/less the item gets endorsed), a correlation of −1 implies the opposite, and a correlation of 0 implies that the perceiver rated the target irrespective of the items' evaluative connotations. In their original study, Kwan et al. (2004) simply keyed all items in the desirable direction and then separately averaged perceiver- and target-effects across items. This procedure is less exact than ours because it gives equal weight to all items, irrespective of how evaluative they are, and thereby introduces error variance (leading to lower power). Also, our approach makes the overall positivity of judgments quantifiable within a common metric (correlation coefficients), and thus directly comparable between studies.

The positivity of the participants' self-judgments was compared with the average positivity of the judgments of the same participants by the four standard perceivers, and with the average positivity of the same participants' judgments of the four standard targets. Again we modified Kwan et al.'s approach somewhat, as we *regressed* the first variable on the latter two variables, and interpreted the *residuals* as SEB, rather than computing simple difference scores (Paulhus & John, 1998). The main advantage of doing so is that the resulting measure of SEB will, by definition, be independent of the two predictors (i.e., the positivities of standard perceiver and standard target judgments), and may thus be independently interpreted. Another advantage is that the approach of using regression residuals is more in keeping with the types of analyses (correlation, regression) that are typically used in this research field, and in the present article.

First we computed the above-described positivity correlations (i.e., profile correlations between item endorsements and item desirabilities across the 46 adjectives) for each description of a target by a perceiver. The positivity of a given participant's average judgment of the four standard targets may be interpreted as that perceiver's evaluative "perceiver-effect," that is, the perceiver's tendency to judge other people positively (cf. Srivastava, Guglielmo, & Beer, 2010; Wood, Harms, & Vazire, 2010). The positivity of the four standard perceivers' average judgment of a given

Table 2. Basic Statistics Regarding Positivity Indices.

Type of judgment	Reliability	Mean (*SD*)	Correlations		
			2	3	4
Self (1)	.80	.66 (.32)	.39**	.03	.92**
Standard perceivers (2)	.92	.46 (.51)		−.13	.00
Standard targets (3)	.84	.60 (.37)			.00
Self-enhancement bias (4)	.77	.00 (.29)			

Note. N = 201. Reliability = split-half–corrected correlation between positivity indices based on two random halves of the item sample; self-enhancement bias = residuals from regressing self-ratings on averaged standard perceiver ratings and averaged standard target ratings.

participant may be interpreted as that participant's evaluative "target-effect," that is, his or her tendency to be judged positively by other people. We computed a total of 201 profile correlations between item desirabilities and self-ratings, 201 profile correlations between item desirabilities and average ratings *of* others (i.e., standard targets), and 201 profile correlations between item desirabilities and average ratings *by* others (i.e., standard perceivers). Then we subjected these correlations to Fisher's r-to-Z transformation and used them in a multiple regression analysis. Specifically, we simultaneously predicted the positivities of the participants' self-judgments from the positivities of their average judgments of and by others. The *residuals* resulting from this regression constitute our measure of SEB (i.e., the tendency to judge oneself more positively than would be expected based on (a) how positively one judges others and (b) how positively one is judged by others).

Results

Basic Statistics Regarding the Positivity Indices

Table 2 displays the Spearman-Brown–corrected split-half reliabilities (using random splits of the 46 adjectives), the means and standard deviations, and the intercorrelations of the various positivity indices. All indices were highly reliable, and several associations between them were noteworthy: First, as in Kwan et al. (2004), evaluative perceiver- and target-effects were not significantly associated with one another ($r = -.13$). That is, judging others more positively was unrelated to being judged more positively by others. Second, the positivity of the participants' self-judgments was moderately related to the positivity of the standard perceiver judgments ($r = .39$). Thus, participants whose behavior was judged more positively by the standard perceivers also tended to judge their own behavior more positively. Third, however, the positivity of the participants' self-judgments was unrelated to the positivity of their judgments of the standard targets ($r = .03$). This implies that, in the present study, evaluative perceiver-effects (tendencies to judge other people more or less positively) could essentially have been ignored in computing the SEB. Fourth, the SEB was perfectly (and inevitably) independent of the positivity of the standard perceiver and standard target ratings, whereas at the same time being highly correlated with the positivity of the participants' self-ratings. This is the ideal correlation pattern one would hope for in a proper index of SEB (see above). All four indices were normally distributed (K-S-tests: $Z < 0.71$, $p > .700$).

The average positivity was .66 ($SD = .32$) for the participants' self-judgments, but only .46 ($SD = .51$) for the judgments of the participants by the standard perceivers, and .60 ($SD = .37$) for the judgments of the standard targets by the participants. A repeated-measures analysis of variance, $F(2, 199) = 17.03$, $p < .001$, revealed that the participants did judge themselves significantly more positively, on average, than they were judged by the four standard perceivers (Cohen's $d = 0.55$; $p < .001$), but not significantly more positively than they judged the four standard targets (Cohen's $d = 0.25$; $p = .094$). Thus, in Kwan et al.'s (2004) terminology, the average participant did display a self-insight bias, but not a social comparison bias.

Predicting Self-Rated Intelligence From Actual Intelligence and Self-Enhancement Bias

Responses to items of self-report questionnaires may be assumed to reflect "the truth" regarding the targets to some extent, as well as the perceivers' tendencies to self-enhance or self-denigrate (McCrae & Costa, 1983). We investigated the relative contributions of truth and bias to self-ratings, using the measurement domain of intelligence. The intelligence domain was chosen for two reasons: First, there is good agreement that people's "actual" intelligence is measurable by means of standardized tests, so our choice of *intelligence tests* as an accuracy criterion should be widely acceptable. Second, intelligence is a highly valued personality trait. In fact, Leising, Ostrovski, and Borkenau (2012) found that among the 758 terms their German research participants came up with when asked to describe themselves and others, "intelligent" was the most positively evaluated of *all* terms. Likewise, in N. Anderson's (1968) study, "intelligent" was the seventh most positively evaluated of

555 English terms. Therefore, self-ratings of intelligence should be strongly susceptible to being affected by people's self-enhancement or self-derogation biases.

Our set of personality questionnaires contained the same item list that was also used for rating the participants' behavior in the lab. Among these items were the terms *kenntnisreich* (*knowledgeable*), and *klug* (*smart*, fourth most positive in Leising et al., 2012). We used the average ($\alpha = .79$) of these two items as the outcome variable in a multiple regression analysis, and tried to predict it from the participants' actual intelligence, as measured by the average of the three intelligence tests, and from the SEB. In this analysis, the SEB was computed omitting the two intelligence items, so predictor and criterion were not contaminated with one another. The resulting model, $F(2, 198) = 15.98$, $R^2 = .14$, $p < .001$, showed that actual intelligence (standardized beta = 0.27, $t = 4.09$, $p < .001$) and the SEB (standardized beta = 0.27, $t = 4.11$, $p < .001$) made significant contributions of the same size in predicting self-rated intelligence. Repeating this analysis for the two intelligence items separately yielded virtually identical results. So, as expected, self-ratings of intelligence *did* reflect both SEB and the truth.

If self-ratings of intelligence are substantially affected by SEB, then SEB might operate as a suppressor variable, and controlling for SEB should lead to stronger associations between self-rated and actual intelligence. However, the standardized betas just reported (and the identical semipartial correlations) were only marginally larger than the respective zero-order correlations between predictors (intelligence, 0.26; SEB, 0.26) and criterion (self-assessed intelligence). Thus, even though the participants' tendency to self-enhance or self-denigrate clearly affected their ratings of their own intelligence, it did not act as a suppressor, because using the SEB as a covariate did not lead to any substantial increase in validity. This finding is well in line with most of the literature (e.g., Paunonen & LeBel, 2012), and will be discussed further below. We also used moderated regression analysis to test for possible moderation effects of SEB on the association between self-rated and actual intelligence. Specifically, we entered the product of actual intelligence and SEB as an additional predictor into the regression. Doing so enabled us to test whether the association between actual and self-rated intelligence was stronger for participants low in SEB (Borkenau & Ostendorf, 1992). However, the standardized beta for the product was not significant (−.05, $p = .455$), whereas the standardized betas for actual intelligence and SEB did not change. Thus, SEB did not function as a moderator either.

Is the Effect of Self-Enhancement Bias on Self-Ratings Moderated by Item Desirability?

For the analysis just presented, we chose a highly evaluative trait (intelligence) as the criterion variable, because that made it likely that self-ratings would be affected by individual differences in SEB, besides actual intelligence. As a next step, we tested the assumption that item desirability moderates the effect of SEB on self-ratings more systematically: First, we separately correlated the participants' SEBs with their responses to each of the 157 items of the personality questionnaires (RSE, NPI, LOT-R, BDI, BIDR, SES-17). This resulted in 157 correlation coefficients reflecting the extent to which the participants' responses to each item could be predicted from their individual SEBs. Second, these correlation coefficients were *r*-to-*Z* transformed and then correlated with the average ratings of the items' social desirability. The stronger this correlation, the better the items' proneness to reflecting self-enhancement or –denigration may be predicted from how much the item has desirable or undesirable content.

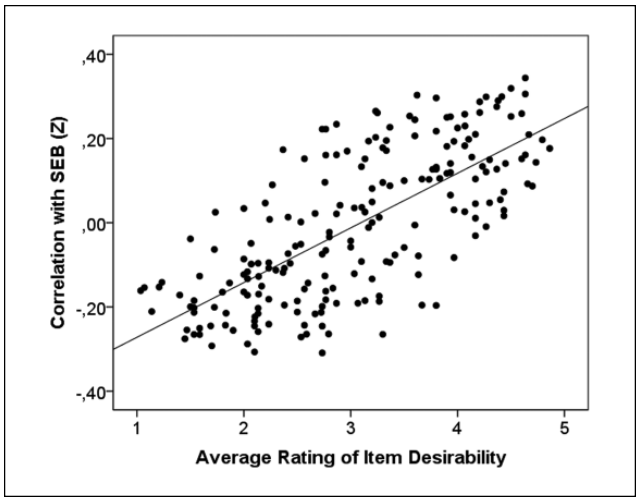

Figure 1. Predicting the (*r*-to-*Z* transformed) associations between personality item endorsement and self-enhancement bias (SEB) from rated item desirability, $r(155) = .64$, $p < .001$.

The results were very clear: The extent to which the participants' responses to the individual questionnaire items could be predicted from SEB ranged from $r = -.32$ to $r = .31$, and (after *r*-to-*Z* transformation) correlated at $r(155) = .64$, $p < .001$ with the rated desirabilities of the items. Accounting for nonindependency by means of a multilevel model led to the exact same conclusion. An inspection of the scatter plot (see Figure 1) revealed that this effect was nicely linear and symmetric: Items with undesirable content showed negative associations with SEB (i.e., self-enhancers rated themselves lower on these items), and items with desirable content showed positive associations (i.e., self-enhancers rated themselves higher on these items). For example, participants with stronger self-enhancement tendencies tended to agree *less* with the Rosenberg item "I certainly feel useless at times" (average desirability rating: 1.59), $r = -.25$, whereas they tended to agree *more* with the Rosenberg item "I take a positive attitude towards myself" (average desirability rating: 4.50),

$r = .25$. Evaluatively neutral items, on the other hand, did not reflect the participants' tendencies to self-enhance or –denigrate to the same extent. For example, responses to the BIDR item "I sometimes drive faster than the speed limit" (average desirability rating: 3.00) correlated at only $r = -.04$ with the participants' SEBs.

Thus, item desirability clearly *did* moderate the influence of SEB on self-ratings of personality (cf. Bäckström et al., 2009; John & Robins, 1993; Leising & Borkenau, 2011; McCrae & Costa, 1983; Vazire, 2010). Note that the highest correlation (.31) was only slightly higher than the one we found when predicting self-ratings of intelligence (see above). This means that the effect of self-enhancement on self-ratings of intelligence was close to the upper limit of what may be expected, which accords well with the fact that intelligence is one of the most positively valued traits.

Associations Between Judgment Positivity Indices and Questionnaire Self-Ratings

As a next step, we investigated the extent to which the participants' self-ratings on the various questionnaire scales were associated with the three different indices of judgment positivity. We performed these analyses not only at the level of the individual self-report scales but also at the level of broader factors, to minimize redundancy among the different constructs as much as possible. When the 157 items were subjected to a joint principal components analysis, the Scree plot (eigenvalues: 17.04, 9.38, 5.82, 4.98, 4.63) clearly suggested the presence of *two* predominant factors, corroborating the outcome of a previous study in which a similar set of measures was factored (Leising, Ostrovski, & Zimmermann, 2013). A couple of weaker, more content-specific factors could also be identified, but we chose to ignore them as they did not make additional contributions to the association between self-ratings and positivity indices.

After Promax rotation, the highest loading items on the first factor (according to the pattern matrix) were "On the whole, I am satisfied with myself" (.72; RSE), "I certainly feel useless at times" (−.70; RSE), "I accuse myself for my mistakes and weaknesses" (−.67; BDI), "All in all, I am inclined to feel that I am a failure" (−.67; RSE), and "I take a positive attitude towards myself" (.66; RSE). The highest loading items on the second factor were "I see myself as a good leader" (.67; NPI), "I would prefer to be a leader" (.67, NPI), "I have a natural talent for influencing people" (.65, NPI), "I have a strong will to power" (.63; NPI), and "I really like to be the center of attention" (.63; NPI).

This factor structure closely resembled the one previously reported by Leising et al. (2013), who factored a similar set of measures. Thus, the interpretation of the two factors was relatively straightforward: The first factor reflects people's overall contentedness with themselves. In the Leising et al. (2013) study, this factor was called "Positive Self-Regard" (PSR), a nomenclature that we will adopt in the present article. However, we believe that this factor is basically identical with the broad general self-evaluation factor that has been identified by several authors, and given various names such as "Demoralization" (Tellegen, 1985; inversely keyed), "Personal Negativity" (Furr & Funder, 1998; inversely keyed), "Core Self-Evaluation" (Bono & Judge, 2003), and "Vulnerability" (Pincus et al., 2009, inversely keyed). The second factor reflects people's subjective capacity and motivation to lead others. It is essentially identical with the main factor of the NPI (Ackerman et al., 2011; Kubarych, Deary, & Austin, 2004). However, Leising et al. (2013) chose to call this factor "Claim to Leadership" (CTL) in order to avoid the widespread conceptual confusion around the term *narcissism* (Cain, Pincus, & Ansell, 2008; Miller & Campbell, 2008). The second factor is also highly akin to the "Grandiosity" factor from the "Pathological Narcissism Inventory" (Pincus et al., 2009; Wright, Lukowitsky, Pincus, & Conroy, 2010). Overall, the outcome of this analysis suggests that the overlap among the constructs that are assessed by the questionnaires we used in our study is considerable, and largely explainable in terms of two common factors. Their factor scores were significantly, but only moderately correlated with one another, $r(199) = .28$, $p < .001$. Complete factor loading matrices, as well as the raw data, are available from the first author, on request.

Table 1 displays the correlations between the individual self-report scales and the two broad factors, as well as the correlations of the scales and factors with the different indices of judgment positivity. As N is 201, correlations whose absolute value exceeds .14 are significant at $p < .05$. The individual scales reflected various mixtures of the two factors: The PSR factor was most clearly marked by high self-esteem (RSE), $r = .84$; low depression (BDI), $r = -.90$; and high optimism (LOT-R), $r = .73$. The correlations between these three scales and the Claim to Leadership factor on the other hand were rather moderate, $r < .40$. Claim to Leadership was most clearly marked by the NPI, $r = .95$, which in turn showed a rather moderate correlation, $r = .37$, with PSR.

All scales and factors showed significant associations with a tendency to portray one's own behavior in the lab positively (fourth data column), with coefficients ranging from $r = .25$ (SES-17) to $r = .51$ (PSR). Likewise, all scales and factors showed significant associations with the SEB (seventh data column), with coefficients ranging from $r = .24$ (BIDR-IM) to $r = .43$ (PSR). Notably, none of the so-called Social Desirability scales (BIDR-IM, BIDR-SDE, SES-17) outperformed the Rosenberg Scale, the BDI, the Life Orientation Test, or the NPI in assessing self-enhancement. Thus, these latter scales were just as valid measures of self-enhancement as were scales designed especially for that purpose.

When the two factors (PSR, CTL) were simultaneously entered as predictors of SEB in a multiple regression analysis, corrected R^2 = .21, $F(1, 198)$ = 27.65, $p < .001$, both PSR (standardized beta = .38, $p < .001$) and CTL (standardized beta = .18, p = .007) independently predicted the SEB. This finding also replicates a previous one (Leising et al., 2013). It is important because it suggests that self-enhancement is associated with two different forms of self-evaluation, and that these effects are largely *additive*: The persons who will show the largest SEB are the ones who (a) have a positive attitude toward themselves (high PSR) *and* (b) consider themselves good leaders (high CTL).

We found a relatively differentiated pattern in regard to how the individual scales were associated with the positivity of judgments by the standard perceivers and judgments of the standard targets. Participants with higher self-esteem (r = .27), greater optimism (r = .37), and lower depression (r = −.24) were judged more positively by the standard perceivers (fifth data column). Obviously, participants who were happier with themselves overall made better impressions on the standard perceivers. In contrast, none of the social desirability scales predicted more positive evaluations by the standard perceivers.

The NPI was the only scale that significantly predicted more negative judgments of the standard targets by the participants (r = −.17). Thus, more "narcissistic" individuals actually had a tendency to denigrate others a bit, which is well in line with a more *interpersonal* interpretation of this scale (Leising et al., 2013). Taken together, it can be said that associations between personality self-ratings and evaluative target-effects (i.e., the positivity of averaged standard perceiver ratings) were generally stronger than associations between personality self-ratings and evaluative perceiver-effects (i.e., the positivity of averaged standard target ratings).

Associations Between Self-Enhancement Bias and Life Satisfaction Ratings

We were also interested in determining the extent to which the participants' ratings of their own life satisfaction across various domains were associated with the SEB. To investigate this, we first factored the participants' judgments of their satisfaction with 21 different aspects of their own lives. Judgments of satisfaction with one's children were omitted, because only 17 participants reported having children. The first factor (eigenvalue = 4.33) was much stronger than any subsequent factors (2.10, 1.86, 1.61, . . .), so for the sake of simplicity we dismissed the latter. There was a significant association between the general life satisfaction factor and the SEB, $r(113)$ = .38, $p < .001$. However, a relatively differentiated picture emerged with regard to associations between specific domains of life satisfaction and the SEB: Seven domains of life satisfaction had significant ($p < .05$) associations with the SEB—work, $r(75)$ = .44; appearance, $r(113)$ = .34; financial situation, $r(113)$ = .23; sex life, $r(113)$ = .24; clothing style, $r(113)$ = .23; parents, $r(112)$ = .28; future prospects, $r(113)$ = .22; all ps < .02—whereas the remaining 13 did not.

Discussion

We presented an empirical investigation based on Kwan et al.'s (2004) conceptualization of self-enhancement as a tendency to judge oneself more positively than one is judged by others and more positively than one judges others. However, our study incorporated two important deviations from the original proposal by Kwan et al. (2004). First, we applied the model to the overall positivity of rating profiles, as measured by the profile correlation between item endorsements and item desirabilities. The advantages of doing so are that (a) items are directly weighted in accordance with how evaluative they are (reducing error variance) and (b) the amount of social desirability involved in judgments becomes directly quantifiable, and thus comparable between studies. Second, rather than using difference scores between self-ratings and ratings of and by others, we regressed the positivity of the self-ratings on the positivity of the other ratings, and interpreted the residuals as the SEB (Paulhus & John, 1998). The advantage of doing so is that the SEB becomes independent of the two predictors, and may thus be independently interpreted. We used this modified version of the Kwan et al. (2004) approach to answer a set of important research questions in regard to socially desirable responding (see page 2f. for the complete list of questions).

We demonstrated that all indices of judgment positivity had good reliability, as random samples of items produced positivity estimates that correlated highly with one another. The positivity of the targets' self-judgments did correlate with the positivity of the standard perceiver judgments, but, unexpectedly, did not correlate with the positivity of the participants' judgments of the standard targets. Thus, participants who judged others more positively did not judge themselves more positively as well. This could be interpreted as evidence that evaluative perceiver-effects may essentially be ignored in computing self-enhancement. However, given that the number of studies investigating these issues with advanced methodology is still very small, we prefer to abstain from making such a broad claim. More research is clearly needed to clarify whether the overall positivity of people's self-images is (un-)related to how positively they view others.

The analysis in which we predicted self-ratings of intelligence from the participants' intelligence test scores and the SEB showed that the latter variable fully conformed to

expectations: It predicted self-ratings of a highly desirable trait (intelligence) independent of actual trait level. This is what one would hope for in a proper measure of self-enhancement. At the same time, however, controlling for the SEB (as a possible suppressor or moderator) did not lead to any notable improvement in the validity of intelligence self-ratings, a finding that is well in line with most of the published literature (Borkenau & Ostendorf, 1992; McCrae & Costa, 1983; Piedmont et al., 2000). Paunonen and LeBel (2012) recently used a simulation approach to demonstrate that such seemingly contradictory findings (contamination of self-ratings with SEB; but failure to improve the validity of self-ratings by controlling for SEB) may be viewed as the default outcome for this kind of studies. One reason is that, in order be able to improve the validity of self-ratings by controlling for SEB, the contamination of the self-ratings with SEB would have to be extremely strong. In our study the contamination rarely exceeded an absolute value of $r = .30$, even for the most evaluative traits, so substantial validity gains by controlling for SEB were unlikely. It seems possible, however, that under circumstances where incentives for self-enhancement are stronger (e.g., personnel selection), contamination—and thus the potential for validity gains—would increase. Future research should address this possibility.

The extent to which the SEB predicted personality self-ratings could very well be predicted from the rated social desirability of the respective items. This finding confirms the assumption that the perceiver's evaluative attitude toward the target (operationalized as SEB in the present study) and the item's evaluative tone closely interact in shaping personality ratings (Bäckström et al., 2009; Leising & Borkenau, 2011; Leising et al., 2010; Leising et al., 2014; McCrae & Costa, 1983; Saucier, 1994). The predictive power of the SEB in these analyses ranged from about $r = -.30$ to $r = .30$. It should be noted, however, that these were the associations we found for single items. When using a higher level of aggregation, individual differences in self-enhancement may be expected to affect participants' responses even more strongly (e.g., the broad PSR factor correlated at $r = .43$ with the SEB).

The shared variance of the 157 questionnaire items could largely be explained in terms of two factors. The first of these (PSR) had strong positive correlations with the Rosenberg Self-Esteem Scale and the Life Orientation Test–Revised (i.e., dispositional optimism), as well as a strong negative correlation with the BDI. The second factor (Claim to Leadership) closely overlapped with the measurement domain of the total score of the NPI. The present research replicated previous studies showing that both factors predict self-enhancement (Farwell & Wohlwend-Lloyd, 1998; Gabriel, Crittelli, & Ee, 1994; John & Robins, 1994). However, the present study is only the second to demonstrate that the two factors *independently* predict self-enhancement (cf. Leising et al., 2013). So the people who will self-enhance the most arethose who consider themselves born leaders (high CTL) *and* are generally happy with themselves (high PSR). In contrast, people who shy away from leadership positions (low CTL) and are unhappy with themselves (low PSR) will tend to underestimate themselves considerably, whereas people who have high scores on one factor but low scores on the other factor will have more realistic self-images (because then the effects of CTL and PSR should cancel out).

Notably, none of the so-called social desirability scales outperformed any of the other scales in predicting the SEB. Thus, the NPI, the LOT-R, the BDI, and the RSE are just as potent—or even more potent—indicators of socially desirable responding as are the BIDR and the SES-17. They all reflect self-enhancement at about $r = .30$ (cf. Table 1). However, the social desirability scales were not systematically associated with the positivity of the standard perceiver judgments, whereas the RSE, the BDI, and the LOT-R were. An optimistic reading of these results would suggest that the social desirability scales actually do differ from the other scales in one important respect: Whereas all scales assess socially desirable responding to a similar extent, the social desirability scales do not also predict actual behavior (at least not in our lab tasks). Thus, the results of the present study may be interpreted as suggesting that the social desirability scales we used assess only style, but not substance. This conclusion, however, hinges on the assumption that our participants' behavior in the lab was sufficiently representative of their behavior in the outside world. Future research will thus have to address the question of whether social desirability scales differ from other scales in regard to their ability to predict real life behavior.

Our analyses focusing on life satisfaction showed that there is a strong general factor accounting for self-rated life satisfaction across various domains, and that this factor is associated with the SEB, $r = .38$. The cross-sectional design of the present study does not permit any analyses of the direction of effects, however. Thus, there are at least two plausible interpretations of these associations: Either, people holding overly positive views of themselves (high SEB) may also tend to see various aspects of their own lives too positively. According to this interpretation, a person's subjective life satisfaction might at least partly reflect a judgment bias: Given the exact same life circumstances, different persons might report quite different levels of life satisfaction, depending on their individual SEB levels. However, it would also be possible that people who have good reasons for being satisfied with their own lives (e.g., they actually *have* better workplaces or sex lives), develop a positively biased self-image (high SEB) as a result. According to this interpretation, the life satisfaction ratings would at least partly reflect a reality, and a more positive

reality might make people see themselves (e.g., their own behavior in the lab) more positively. The present study does not enable a decision as to which of these two explanations is more valid.

In interpreting the findings of the present study, we also need to consider the possibility that the participants' ratings of their own behavior in the lab may reflect other influences apart from (a) their actual behavior and (b) their tendencies to view themselves more positively or negatively. In fact, studies have shown that peoples' ratings of their own behavior do reflect their general self-images (e.g., as "quick" or "lazy" in particular) to some extent, even if their "actual" behavior in the situation (as judged by observers at zero acquaintance) is controlled for (Leising, 2011; Leising et al., 2014; Sadler & Woody, 2003). Such content-specific "consistency biases" may also have affected the participants' self-ratings in the present study. However, as we used a broad set of adjectives covering many different content domains, and profile correlations that quantify the overall positivity of personality profiles *across* items, it may be assumed that such biases merely constituted unsystematic measurement error in the end.

Summary and Conclusion

The present article established a new measure of socially desirable responding based on the reasoning of Kwan et al. (2004). We introduced two major refinements of their approach, to make the measure even better interpretable. Using this new measure, we showed that participants' responses to self-report questionnaires do reflect SEB, and that they do so the more the respective items have evaluative (i.e., positive or negative) content. Notably, scales designed particularly for the purpose of assessing socially desirable responding (SES-17, BIDR) did not reflect SEB more than did the other scales (RSE, BDI, etc.). Also, no suppressor or moderator effects of socially desirable responding on the validity of self-ratings of intelligence were found. Taken together, the study demonstrates that socially desirable responding is measurable and does affect self-reports of personality substantially. At the same time, it demonstrates that so-called "lie" or "social desirability" scales lack discriminant validity as compared with other self-evaluative personality scales.

Appendix

List of Adjectives

The Borkenau and Ostendorf (1998) List. Temperamentvoll (Vivacious), Launisch (Erratic), Rücksichtsvoll (Considerate), Gefühlsstabil (Emotionally Stable), Kenntnisreich (Knowledgeable), Scheu (Shy), Egoistisch (Egoistical), Kontaktfreudig (Sociable), Konsequent (Consistent), Klug (Smart), Schweigsam (Taciturn), Einfallslos (Unimaginative), Gutmütig (Good-Natured), Unbeständig (Unsteady), Verletzbar (Vulnerable), Dynamisch (Dynamic), Unkundig (Uninformed), Rechthaberisch (Dogmatic), Fleißig (Industrious), Empfindlich (Touchy), Verantwortungsbewusst (Responsible), Arbeitsscheu (Work-Shy), Leichtsinnig (Reckless), Hilfsbereit (Ready to Help), Geistreich (Witty), Gelassen (Relaxed), Unempfindlich (Robust), Phantasielos (Fanciless), Herrschsüchtig (Dictatorial), Zurückhaltend (Reserved).

From the Interpersonal Adjective List (Jacobs & Scholl, 2005). Durchsetzungsfähig (Assertive), Zynisch (Cynical), Gehorsam (Obedient), Provokativ (Provocative), Aufgeschlossen (Open-Minded), Boshaft (Malicious), Ungesellig (Unsociable), Schüchtern (Bashful), Feindselig (Hostile), Selbstsicher (Self-assured), Einfühlsam (Empathetic), Verschlossen (Tight-Lipped), Herzlich (Cordial), Still (Silent), Folgsam (Compliant), Kommunikativ (Communicative).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author received a grant from the German Research Foundation (DFG). The grant number was LE 2151/3-1.

References

Ackerman, R. A., Witt, E. A., Donnellan, M. B., Trzesniewski, K. H., Robins, R. W., & Kashy, D. A. (2011). What does the Narcissistic Personality Inventory really measure? *Assessment*, *18*, 67-87. doi:10.1177/1073191110382845

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Anderson, J. L., Sellbom, M., Wygant, D., & Edens, J. F. (2013). Examining the necessity for and utility of the Psychopathic Personality Inventory–Revised (PPI-R) validity scales. *Law and Human Behavior*, *37*, 312-320. doi:10.1037/lhb0000018

Anderson, N. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *9*, 272-279. doi:10.1037/h0025907

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

Bäckström, M., Björklund, F., & Larsson, M. R. (2014). Criterion validity is maintained when items are evaluatively neutralized: Evidence from a full scale five-factor model inventory. *European Journal of Personality*, *28*, 620-633. doi:10.1002/per.1960

Baity, M. R., Siefert, C. J., Chambers, A., & Blais, M. A. (2007). Deceptiveness on the PAI: A study of naïve faking with psychiatric inpatients. *Journal of Personality Assessment*, *88*, 16-24. doi:10.1080/00223890709336830

Beck, A. T., & Steer, R. A. (1987). *Beck Depression Inventory (BDI)*. San Antonio, TX: Psychological Corporation.

Bono, J. E., & Judge, T. A. (2003). Core self-evaluations: A review of the trait and its role in job satisfaction and job performance. *European Journal of Personality*, *17*, S5-S18. doi:10.1002/per.481

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, *86*, 599-614. doi:10.1037/0022-3514.86.4.599

Borkenau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality*, *6*, 199-214. doi:10.1002/per.2410060303

Borkenau, P., & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual variations over time? *Journal of Research in Personality*, *32*, 202-221.

Cain, N. M., Pincus, A. L., & Ansell, E. B. (2008). Narcissism at the crossroads: Phenotypic description of pathological narcissism across clinical theory, social/personality psychology, and psychiatric diagnosis. *Clinical Psychology Review*, *28*, 638-656. doi:10.1016/j.cpr.2007.09.006

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston, MA: Houghton Mifflin.

Collani, G., & Herzberg, P. Y. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg [A revised version of the German adaptation of Rosenberg's self-esteem scale]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *24*, 3-7. doi:10.1024//0170-1789.24.1.3

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

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37*, 90-93. doi:10.1037/h0058073

Farwell, L., & Wohlwend-Lloyd, R. (1998). Narcissistic processes: Optimistic expectations, favorable self-evaluations, and self-enhancing attributions. *Journal of Personality*, *66*, 65-83. doi:10.1111/1467-6494.00003

Ferring, D., & Filipp, S.-H. (1996). Messung des Selbstwertgefühls: Befunde zu Reliabilität, Validität und Stabilität der Rosenberg Skala [Measurement of self-esteem: Findings on reliability, validity, and stability of the Rosenberg scale]. *Diagnostica*, *42*, 284-292.

Furr, R. M., & Funder, D. C. (1998). A multimodal analysis of personal negativity. *Journal of Personality and Social Psychology*, *74*, 1580-1591. doi:10.1037//0022-3514.74.6.1580

Gabriel, M. T., Critelli, J., & Ee, J. S. (1994). Narcissistic illusions in self-evaluations of intelligence and attractiveness. *Journal of Personality*, *62*, 143-155. doi:10.1111/j.1467-6494.1994.tb00798.x

Glaesmer, H., Hoyer, J., Klotsche, J., & Herzberg, P. Y. (2008). Die deutsche Version des Life-Orientation-Tests (LOT-R) zum dispositionellen Optimismus [The German version of the Life-Orientation-Test (LOT-R) for dispositional optimism and pessimism]. *Zeitschrift für Gesundheitspsychologie*, *16*, 26-31. doi:10.1026/0943-8149.16.1.26

Hofstee, W. K. B., & Hendriks, A. A. J. (1998). The use of scores anchored at the scale midpoint in reporting individuals' traits. *European Journal of Personality*, *12*, 219-228. doi:10.1002/(SICI)1099-0984(199805/06)12:3<219::AID-PER315>3.0.CO;2-Y

Horn, W. (1983). *Leistungsprüfsystem. Handanweisung* [Performance test system, 2nd ed.]. Göttingen, Germany: Hogrefe.

Jacobs, I., & Scholl, W. (2005). Interpersonale Adjektivliste (IAL). Die empirische Umsetzung theoretischer Circumplex-Eigenschaften für die Messung interpersonaler Stile [Interpersonal Adjective List; IAL: An empirical implementation of theoretical circumplex concepts for the assessment of interpersonal style]. *Diagnostica*, *51*, 145-155. doi:10.1026/0012-1924.51.3.145

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, *61*, 521-551. doi:10.1111/j.1467-6494.1993.tb00781.x

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, *66*, 206-219. doi:10.1037//0022-3514.66.1.206

Kenny, D. A. (1994). *Interpersonal perception. A social relations analysis*. New York, NY: Guilford Press.

Kubarych, T. S., Deary, I. J., & Austin, E. J. (2004). The Narcissistic Personality Inventory: Factor structure in a non-clinical sample. *Personality and Individual Differences*, *36*, 857-872. doi:10.1016/S0191-8869(03)00158-2

Kwan, V. S. Y., John, O. P., Kenny, D. A., Bond, M. H., & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review*, *111*, 94-110. doi:10.1037/0033-295X.111.1.94

Lehrl, S. (2005). *Mehrfachwahl-Wortschatz-Intelligenztest* [Multiple-choice vocabulary intelligence test]. Balingen, Germany: Spitta.

Leising, D. (2011). The consistency bias in judgments of one's own interpersonal behavior. Two possible sources. *Journal of Individual Differences*, *32*, 137-143. doi:10.1027/1614-0001/a000046

Leising, D., & Borkenau, P. (2011). Person perception, dispositional inferences and social judgment. In L. M. Horowitz & S. Strack (Eds.), *Handbook of interpersonal psychology: Theory, research, assessment, and therapeutic interventions* (pp. 157-170). Hoboken, NJ: Wiley.

Leising, D., Borkenau, P., Zimmermann, J., Roski, C., Leonhardt, A., & Schütz, A. (2013). Positive self-regard and claim to leadership: Two fundamental forms of self-evaluation. *European Journal of Personality*, *27*, 565-579. doi:10.1002/per.1924

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology*, *98*, 668-682. doi:10.1037/a0018771

Leising, D., Gallrein, A.-M. B., & Dufner, M. (2014). Judging the behavior of people we know: Objective assessment, confirmation of preexisting views, or both? *Personality and Social Psychology Bulletin*, *40*, 153-163. doi:10.1177/0146167213507287

Leising, D., Ostrovski, O., & Borkenau, P. (2012). Vocabulary for describing disliked persons is more differentiated than vocabulary for describing liked persons. *Journal of Research in Personality*, *46*, 393-396. doi:10.1016/j.jrp.2012.03.006

Leising, D., Ostrovski, O., & Zimmermann, J. (2013). Are we talking about the same person here? Inter-rater agreement in judgments of personality varies dramatically with how much the perceivers like the targets. *Social Psychological and Personality Science*, *4*, 468-474. doi:10.1177/1948550612462414

Locke, K. D., & Horowitz, L. M. (1997). The multifaceted self effect: Flexibility or merely self-enhancement? *Journal of Research in Personality*, *31*, 406-422.

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, *51*, 882-888. doi:10.1037/0022-006X.51.6.882

McGrath, R. E., Kim, B. H., & Hough, L. (2011). Our main conclusion stands: Reply to Rohling et al. (2011). *Psychological Bulletin*, *137*, 713-715. doi:10.1037/a0023645

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

Miller, J. D., & Campbell, W. K. (2008). Comparing clinical and social-personality conceptualizations of narcissism. *Journal of Personality*, *76*, 449-476. doi:10.1111/j.1467-6494.2008.00492.x

Morey, L. C., & Lanier, V. W. (1998). Operating characteristics of six response distortion indicators for the Personality Assessment Inventory. *Assessment*, *5*, 203-214. doi:10.1177/107319119800500301

Morey, L. C. (2012). Detection of response bias in applied assessment: Comment on McGrath et al. (2010). *Psychological Injury and Law*, *5*, 153-161. doi:10.1007/s12207-012-9131-x

Musch, J., Brockhaus, R., & Bröder, A. (2002). Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit [An inventory for the assessment of two factors of social desirability]. *Diagnostica*, *48*, 121-129.

Oswald, W. D., & Roth, E. (1987). *Zahlen-Verbindungs-Test* [Digit linking test] (2nd ed.). Göttingen, Germany: Hogrefe.

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

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

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

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*, 1025-1060. doi:10.1111/1467-6494.00041

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

Peabody, D., & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology*, *57*, 552-567. doi:10.1037//0022-3514.57.3.552

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, *78*, 582-593. doi:10.1037//0022-3514.78.3.582

Pincus, A. L., Ansell, E. B., Pimentel, C. A., Cain, N. M., Wright, A. G. C., & Levy, K. N. (2009). Initial construction and validation of the Pathological Narcissism Inventory. *Psychological Assessment*, *21*, 365-379. doi:10.1037/a0016530

Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports*, *45*, 590.

Raskin, R. N., & Terry, H. (1988). A principle components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, *54*, 890-902. doi:10.1037//0022-3514.54.5.890

Rohling, M. L., Larrabee, G. J., Greiffenstein, M. F., Ben-Porath, Y. S., Lees-Haley, P., Green, P., & Greve, K. W. (2011). A misleading review of response bias: Comment on McGrath, Mitchell, Kim, and Hough (2010). *Psychological Bulletin*, *137*, 708-712. doi:10.1037/a0023327

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology*, *47*, 213-215. doi:10.1037/0022-006X.47.1.213

Sadler, P., & Woody, E. (2003). Is who you are who you're talking to? Interpersonal style and complementarity in mixed-sex interactions. *Journal of Personality and Social Psychology*, *84*, 80-96. doi:10.1037/0022-3514.84.1.80

Saucier, G. (1994). Separating description and evaluation in the structure of personality attributes. *Journal of Personality and Social Psychology*, *66*, 141-154.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem). A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, *67*, 1063-1078. doi:10.1037//0022-3514.67.6.1063

Schmitt, M., & Maes, J. (2000). Vorschlag zur Vereinfachung des Beck-Depressions-Inventars (BDI) [Simplification of the Beck Depresssion Inventory (BDI)]. *Diagnostica*, *46*, 38-46. doi:10.1026//0012-1924.46.1.38

Schütz, A., Marcus, B., & Sellin, I. (2004). Die Messung von Narzissmus als Persönlichkeitskonstrukt: Psychometrische Eigenschaften einer Lang- und einer Kurzform des Deutschen NPI (Narcissistic Personality Inventory) [Measuring narcissism as a personality construct: Psychometric properties of a long and a short version of the German Narcissistic Personality Inventory]. *Diagnostica*, *50*, 202-218. doi:10.1026/0012-1924.50.4.202

Srivastava, S., Guglielmo, S., & Beer, J. S. (2010). Perceiving others' personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology*, *98*, 520-534. doi:10.1037/a0017057

Stoeber, 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 findings on reliability and validity]. *Diagnostica*, *45*, 173-177.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma & J. D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 681-706). Hillsdale, NJ: Lawrence Erlbaum.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: University of Chicago Press.

Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, *40*, 472-481. doi:10.1016/j.jrp.2005.03.003

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*, 281-300. doi:10.1037/a0017908

Vernon, P. A. (1993). The Zahlen-Verbindungs-Test and other trail making correlates of general intelligence. *Personality and Individual Differences*, *14*, 35-40. doi:10.1016/0191-8869(93)90172-Y

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, *37*, 395-412. doi:10.1037/0022-3514.37.3.395

Wood, D., Harms, P., & Vazire, S. (2010). Perceiver effects as projective tests: What your perceptions of others say about you. *Journal of Personality and Social Psychology*, *99*, 174-190. doi:10.1037/a0019390

Wright, A. G. C., Lukowitsky, M. R., Pincus, A. L., & Conroy, D. E. (2010). The higher order factor structure and gender invariance of the Pathological Narcissism Inventory. *Assessment*, *17*, 467-483. doi:10.1177/1073191110373227