Received: 4 February 2021 | Revised: 13 July 2021 | Accepted: 13 July 2021

DOI: 10.1111/jopy.12662

ORIGINAL MANUSCRIPT

WILEY

Are social desirability scales desirable? A meta-analytic test of the validity of social desirability scales in the context of prosocial behavior

Lukas Lanz[1] | **Isabel Thielmann**[2] | **Fabiola H. Gerpott**[1,3]

[1]WHU – Otto Beisheim School of Management, Düsseldorf, Germany

[2]University of Koblenz-Landau, Landau, Germany

[3]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence
Lukas Lanz, WHU – Otto Beisheim School of Management, Erkrather Straße 224a, 40233 Düsseldorf, Germany.
Email: lukas.lanz@whu.edu

Funding information
The work reported herein was supported by a grant (TH 2318/1-1) from the German Research Foundation obtained by the second author.

Abstract

Social desirability (SD) scales have been used for decades in psychology and beyond. These scales are sought to measure individuals' tendencies to present themselves overly positive in self-reports, thus allowing to control for SD biases. However, research increasingly questions the validity of SD scales, proposing that SD scales measure substantive trait characteristics rather than response bias. To provide a large-scale empirical test of the validity of SD scales, we conducted a meta-analysis ($k = 41$; $N = 8980$) on the relation between SD scale scores and prosocial behavior in economic games (where acting in a prosocial manner is highly socially desirable). If SD scales measure what they are supposed to (namely, SD bias), they should be negatively linked to prosocial behavior; if SD scales measure socially desirable traits, they should be positively linked to prosocial behavior. Unlike both possibilities, the meta-analytic correlation between SD scores and prosocial behavior was close to zero, suggesting that SD scales neither clearly measure bias nor substantive traits. This conclusion was also supported by moderation analyses considering differences in the implementation of games and the SD scales used. The results further question the validity of SD scales with the implication that scholars and practitioners should refrain from using them.

KEYWORDS

economic games, meta-analysis, prosocial behavior, social desirability, social desirability scales

1 | INTRODUCTION

Socially desirable responding (SDR)—the tendency to present oneself overly positive and downplay negative attributes—has frequently been discussed as an issue hampering the validity and interpretation of self-reports (e.g., Arthur et al., 2021; Brownback & Novotny, 2018; Lee & Sargeant, 2011; Ones et al., 1996). A prominent approach that has been proposed to mitigate SDR is the inclusion of social desirability (SD) scales in survey research. The underlying idea is that overly positive self-presentation can be measured and controlled for, thus allowing to obtain participants' "true" (i.e., accurate) levels on socially desirable tendencies and behaviors when using self-reports. SD scales have been—and are still—used by many scholars across disciplines. To illustrate, SD scales were applied to control for response distortion in fields such as public health, medicine, criminology,

and politics (e.g., Hebert et al., 1997; Ng et al., 2020; Vecina et al., 2016; Williams et al., 2009). Even more so than in other fields, SD scales are prominently featured in psychology. For example, SD scales have been used in industrial-organizational psychology (e.g., Arthur et al., 2021), clinical psychology (Perinelli & Gremigni, 2016), social psychology (e.g., Feldman et al., 2017), neuropsychology (e.g., Rodrigues et al., 2015), and personality research (e.g., Liu & Liu, 2021).

Despite their widespread application, there is an ongoing debate (e.g., Connelly & Chang, 2016; Tourangeau & Yan, 2007) about what SD scales actually measure. Whereas some propose that SD scales capture SDR (i.e., a response "style"), others argue that they assess socially desirable personality traits (i.e., "substance" that drive responses). The style versus substance debate is crucial for the understanding and theoretical underpinning of SD as a construct. Even more, the debate is practically relevant for the use of SD scales and the interpretation of corresponding scores in surveys in research and practice. To illustrate, imagine a researcher who includes items from an SD scale in their survey for either measuring style or substance. Depending on which interpretation of the SD scale the researcher relies on, the conclusions drawn from the same SD score may be very different. If the researcher interprets the SD score according to the style perspective and the respondent receives a high score, this would imply that they were "faking good" (i.e., presenting themselves in an overly positive light). By contrast, interpreting the score according to the substance perspective implies that high SD scores indicate socially desirable characteristics, which may support the general reliability of the participant's responses. To conclude, the exact same responses on an SD scale may lead to very different interpretations depending on which meaning of an SD score one adopts. Therefore, it is essential to examine what exactly SD scales measure—and what they don't.

In addition to single studies that contributed to the style versus substance debate (e.g., de Vries et al., 2014; Uziel, 2010, 2014), a meta-analysis (Connelly & Chang, 2016) provided evidence for a blend of both interpretations, implying that SD scores contain method variance (i.e., style) but also—and even more so—trait variance (i.e., substance). However, the scope of this meta-analysis was limited to only one SD scale, the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1991), and relied solely on questionnaire data as a validity criterion, which again might be subject to SDR. More specifically, the authors used discrepancies between personality self- and informant reports as a measure to disentangle method variance (style) from trait variance (substance) in the self-reports. However, this method may not necessarily be unbiased given the imperfect validity of informant reports: Evidence suggests that, much like self-reports, informant reports are often positively biased because informants usually have a close (positive) relationship with the to-be-rated target person and, thus, want to portray the target in a positive light (Leising et al., 2010). The overlap between self- and informant reports may therefore be inflated due to the raters' overly positive self-rating and the informants' tendencies for overly positive other ratings. In turn, even if informant ratings of a target's personality were unbiased, informants may be less accurate in assessing some traits and corresponding behaviors, such as (dis-)honesty (Thielmann et al., 2017). Taken together, even though Connelly and Chang's (2016) study heated up the discussion on the (limited) validity of SD scales, their results may suffer from certain limitations, ultimately calling for additional evidence.

Here, we use an extended and different approach to investigate the validity of SD scales that overcomes these limitations. First, we do not only investigate the BIDR but rely on several prominently featured SD scales and, thereby, broaden the scope of the previous meta-analysis. Further, we rely on a validity criterion that goes beyond reports of behavioral tendencies (as captured in self- and/or informant reports of personality traits) by meta-analytically studying the link of self-reported SD scores with prosocial *behavior*. Prosocial behavior is highly socially desirable because it increases others' welfare and thereby contributes to the functioning of society at large (Platow, 1994; Roussel, 2017). A unique possibility to measure prosocial behavior in controlled experimental settings is provided by *economic games*, which have been established as a standard method across fields (Murnighan & Wang, 2016; Thielmann et al., 2021; van Dijk & Dreu, 2021). These games model different classes of social situations representative of specific real-life situations affording the expression of prosocial behavior. For example, in the widely used dictator game (Forsythe et al., 1994), one individual (the dictator) can decide how to distribute a given (monetary) endowment between themselves and another individual (the recipient). Thus, the game models a social situation in which one individual has full power over a resource that they can use to profit another at personal costs, such as in the case of donations. In turn, the more the dictator gives to the recipient, the more prosocial—and thus socially desirable—is the dictator's action because the recipient directly profits from higher giving. This example illustrates the logic of economic games used to study interpersonal interaction: Individuals are asked to interact in social situations in which their behavior may profit either themselves (selfish behavior) or others (prosocial behavior), with the latter being inherently socially desirable. Thus, these games are well-suited to study the validity of SD scales. Similar to the substance versus style debate of SD scales, recent meta-analytic evidence suggests that prosocial behavior, too, has both trait-like and state aspects (Balliet & van Lange, 2013; Thielmann et al., 2020).

In the current meta-analysis, we relied on six prominent economic games as measures of prosocial behavior. Specifically, we included the (1) dictator game, (2) ultimatum

game, and (3) trust game, as well as the social dilemmas (4) public goods game, (5) prisoner's dilemma game, and (6) commons dilemma (see Table S1 for details on the rules of these games). Although the games differ in terms of the social situation individuals find themselves in and the motives that can guide behavior (Thielmann et al., 2015), all of them provide straightforward measures of prosocial behavior in a specific type of interdependent situation in which own and others' outcomes are intertwined (Thielmann et al., 2021). Overall, utilizing prosocial behavior as expressed in games to validate SD scales prevents that common method variance (i.e., reports of own and others' tendencies and/or behaviors) systematically increases the overlap between SD scales and the validity criterion.

Our meta-analysis provides three key contributions. First, we use the meta-analytic correlation between SD scores and prosocial behavior as a vehicle to test the criterion validity of SD scales. As such, our analysis is based on the logic that studying the link of SD scores with socially desirable *behavior* offers the opportunity to contribute to the understanding of SD scales as capturing style or substance. Second, we seek to get to the essence of the style versus substance debate by additionally studying potential moderators of the relation between SD scales and prosocial behavior. More specifically, we theoretically argue and empirically test the assumption that, if the style interpretation holds, the presence of behavior-contingent incentives and/or complete anonymity in economic games should reduce the correlation between SD scores and prosocial behavior. Third, we test the potential moderating effect of different SD scales used to measure SDR on the relation between SD scores and prosocial behavior, thereby also providing a methodological contribution.

1.1 | SDR and social desirability scales

When individuals are asked to provide a self-report about a certain tendency or behavior, they arguably often consider what would be the socially acceptable—that is, socially desirable—response (Tourangeau & Yan, 2007). While SDR is a commonly used concept, there have been extensive discussions about the essence of SDR and its underlying sub-dimensions. The oldest and arguably most prominent conceptualization of SDR differentiates an Alpha factor capturing social desirability and a Gamma factor capturing self-deceptive enhancement (Wiggins, 1964). Other approaches, in turn, coined the terms self- and other deception to describe differences in the addressee of the response behavior (Sackeim & Gur, 1979), or differentiated between an egoistic and moralistic bias of SDR (Paulhus & John, 1998). In an attempt to combine these two approaches, Paulhus (2002) proposed a further separation of the egoistic bias into self-deceptive enhancement and agency management (i.e., deliberate promotion of competence, fearlessness, physical prowess, etc.), whereas the moralistic bias should be split into self-deceptive denial and communion management (i.e., deliberate minimization of faults). Yet, in essence, SDR can be separated along with its two drivers: gaining a positive reputation and maintaining a positive self-image, as captured in the prominent distinction of IM versus SDE (Paulhus, 1991).

First, SDR signals compliance with a predominant social norm (Tourangeau & Yan, 2007) and should thus be rewarded with social recognition. This instance of SDR primarily focused on others as the recipients of SDR and is covered by the Gamma factor (Wiggins, 1964), other-deception and the more recently introduced moralistic bias (Paulhus & John, 1998). For example, if a person is asked about whether they would help someone in need, this person may agree simply because it is the socially desirable thing to do. Indeed, behaving in a socially desirable manner (or at least claiming to do so) likely has positive consequences in interactions with *others*. Conversely, admitting to behaving in a socially *un*desirable way (e.g., refusing to help someone in need) can result in negative evaluation by others and thus harm oneself. Therefore, respondents should be inherently motivated to present themselves in a positive light to build or maintain a positive reputation. This is particularly true when it is only about one's *self-reported* desirable behavior because the risk of being exposed as a liar is relatively low.

Second, individuals prefer to maintain a positive *self*-image (Paulhus, 1984), meaning that most people want to think of themselves as prosocial and moral beings (Aronson, 1969; Harris et al., 1976). This second instance of SDR is primarily directed at the self and is covered by the Alpha factor, self-deception or the egoistic bias. According to self-maintenance theory (Mazar et al., 2008), individuals might only behave in an immoral (or socially undesirable) way to the extent that they can maintain their positive self-view. In line with this idea, evidence suggests that individuals may only lie to a limited extent (Shalvi et al., 2011, 2015). In combination, both factors—that is, building a positive reputation and maintaining a positive self-image—may pose a threat to the interpretability of survey responses on socially desirable matters.

To measure and control for SDR, scholars have developed different self-report SD scales. For example, the Edwards SD scale (Edwards, 1957), the Minnesota Multiphasic Personality Inventory (MMPI, Butcher et al., 1989), the Marlowe Crowne (MC) scale (Crowne & Marlowe, 1960) and the BIDR (Paulhus, 1991) constitute widely used measures (for an overview of prevalent scales see Table 1). Items of such SD scales describe socially desirable but statistically infrequent behaviors such as “I don't gossip about other people's business” or “I always obey laws, even if I am unlikely to get caught” (Paulhus, 1991). The idea is that individuals who agree with such statements lie because behaving as described is virtually impossible or at least very

TABLE 1 Overview of prominent social desirability scales

Scale	Reference	# items	Response options	IM/SDE separation	*k*
BIDR	Paulhus (1991)	40	Continuous	Yes	11
Eysenck Lie Scale	Eysenck and Eysenck (1964)	6	Dichotomous	No	1
Marlowe Crowne Social Desirability Scale	Crowne and Marlowe (1960)	33	Dichotomous	No	13
SDS-17	Stöber (2001)	16	Dichotomous	No	8
Self-Deception Questionnaire	Sackeim and Gur (1979)	20	Continuous	No	2

Abbreviations: BIDR, balanced inventory of desirable responding; *k*, number of independent samples included in the current meta-analysis that used a respective scale (one study did not include any information about the used scale); SDS-17, social desirability scale 17; # items, number of items.

unlikely. Individuals' scores on SD scales are thus interpreted as indicating how (overly) positive they want to present themselves.

Although different SD scales all intend to measure SDR, they, at least partially, correspond to the different sub-dimensions of SDR. While scales such as the Edwards SD scale and self-deceptive enhancement (SDE) have been shown to load more on the Alpha factor, and others such as the MMPI or impression management (IM) load strongly on the Gamma factor, there the widely used MC scale partially loads on both (Bensch et al., 2019). To illustrate this separation, take the BIDR that differentiates between two sub-dimensions. Both dimensions relate to the motivation to present oneself overly positive, but they target different audiences, namely others versus the self. Accordingly, the sub-dimensions are labeled as *impression management* (IM) and *self-deceptive enhancement* (SDE; Paulhus, 1984). IM describes a deliberate attempt to positively manipulate one's own impression on *others*, for example, by overstating desirable and understating undesirable tendencies and behaviors (Paulhus, 1991). SDE, in turn, describes unconscious *self*-deceit that is aimed at upholding a positive self-image (Paulhus, 1991; Paulhus & John, 1998). Yet, albeit the differentiation in the two sub-dimensions of SD is conceptually well established, it is important to note that in practice, many of the used scales either capture SDR as a unidimensional construct or the researchers using the scales treat the SD scores as such.

Meta-analyses increasingly consider the moderating role of the constructs' operationalizations (e.g., Parks-Leduc et al., 2015; Sibley & Duckitt, 2008; Steel et al., 2008) because different operationalizations of allegedly the same construct may not necessarily provide equivalent measures. Recognizing that many SD scales exist but not all separate the sub-dimension of SDR, we thus test whether the relation between SD scores and prosocial behavior differs across SD scales and across sub-dimensions of IM and SDE. If the type of SD scale moderates the relation between SD scores and prosocial behavior, this would pose new theoretical questions on whether some SD scales are more aligned with a substance or a style interpretation than others.

1.2 | The style versus substance debate

Interpreting SDR in terms of a response style implies that scholars should control for it whenever relying on self-reports. This view is the fundamental assumption of the early SD literature (Crowne & Marlowe, 1960). Until today, the style interpretation of SDR is common among researchers and practitioners who rely on SD scales to control for "faking good" tendencies (Goffin & Christiansen, 2003; Holtgraves, 2004). In support of the style interpretation, evidence suggests that certain SD scales can indeed detect faking. Specifically, some evidence suggests that the Marlowe-Crowne SD scale can serve this purpose (Franzen & Mader, 2019) more so than the BIDR (Lambert et al., 2016). Other studies have, by contrast, suggested that only the BIDR can detect faking, whereas other SD scales, including the Marlowe-Crowne SD scale, cannot (Bensch et al., 2019). In any case, the fact that SD scales are often referred to as "lie scales" (Butcher et al., 1989; Eysenck & Eysenck, 1964; Feldman, 2019; Hathaway & McKinley, 1951) underscores the widespread acceptance of the style interpretation of SD scores.

Following this logic, the style interpretation of SD scales implies that SD scores should be negatively correlated with prosocial behavior. This is because individuals scoring high on SD scales should systematically over*report* their desirable attributes while not behaving accordingly. In other words, regardless of individuals' absolute level of prosociality, high SD scorers should report an even higher level of prosocial (and thus socially desirable) *behavior* than what is actually observed in terms of the behavior individuals show.[1] In line with this reasoning, Paulhus (1984) found that individuals with high SD scores distorted their self-presentation in a positive way to strengthen their image as someone who upholds social norms. Low SD scores, by contrast, should indicate no overreporting, which does not necessarily mean that the respondents have to be honest or virtuous. That is, individuals scoring low on SD scales should simply behave in line with how they report to be, whether it be prosocial or selfish. Taken together, the style interpretation thus implies an overall negative correlation between SD scores and prosocial

behavior that is mainly driven by high SD scores being linked to overreporting one's prosociality.

The validity of SD scales as faking detectors has, however, been heavily criticized in recent years (e.g., Connelly & Chang, 2016; Holtrop et al., 2020; Uziel, 2010; Ward & King, 2018). For example, scholars (e.g., de Vries et al., 2014; Uziel, 2010) argued that SD scores reflect substantive (socially desirable) traits rather than a general response style. The key limitation of SD scales is that it is impossible to differentiate between truly honest respondents who have the (virtuous) traits they claim to have and dishonest respondents who actively lie to present themselves in an overly positive fashion (Tourangeau & Yan, 2007). By implication, if the substance interpretation holds, SD scores should be positively correlated with prosocial behavior: Individuals scoring high on SD scales should behave as reported, that is, in a prosocial manner.

Nonetheless, even if SD scales are interpreted as measures of substantive traits, the question remains which traits SD scales measure. According to one interpretation—particularly referring to the IM sub-dimension—SD scores can be considered as indicators of dispositional *self-control*. This notion is rooted in the finding that IM predicted how well individuals can control themselves and their behavior in social interactions (Uziel, 2010, 2014). Acting in a controlled manner is seen as highly desirable and will, in turn, be rewarded with respect and acceptance by others. Another interpretation of SD scales is based on research linking SD scores to traits reflecting true virtue, such as *honesty-humility* (de Vries et al., 2014). Specifically, individuals *high* in honesty-humility also tend to score higher on SD scales (IM). This finding contradicts the view that SD scales measure lying and instead suggests the exact opposite. Even more conclusive evidence in this regard showed a *negative* relation between SD scores (IM) and actual dishonest behavior in cheating paradigms (Zettler et al., 2015).

Finally, a null correlation between SD scores and prosocial behavior might suggest a mix of both interpretations (i.e., style and substance), which would severely undermine the usability of SD scales to measure either only substance or only style. For example, Müller and Moshagen (2019) asked their participants to play both a hypothetical and an incentivized dictator game. Interestingly, SD scores (IM) accounted for the discrepancy in behavior between games: higher SD scores were positively related to the difference in giving between the incentivized and the hypothetical scenario, which arguably indicates SDR. At the same time, however, SD scores were negatively related to dishonest behavior in a cheating task. This evidence is in line with the meta-analysis by Connelly and Chang (2016), which likewise suggests that SDR is a blend of both style and substance.

Unlike the "blend" interpretation, a null correlation between SD scores and prosocial behavior might also suggest that SD and prosocial behavior are conceptually unrelated. However, this implication is highly unlikely given that (a) prosocial behavior is clearly socially desirable (Platow, 1994; Roussel, 2017) and (b) several items of commonly used SD scales directly refer to prosociality, for example, "I never hesitate to help someone in case of emergency" (Stöber, 2001) or "I never take things that don't belong to me." (Paulhus, 1991). Thus, we maintain that a null relation between SD scores and prosocial behavior arguably supports that SD scales measure a blend of substance and style.

1.3 | The influence of incentives and anonymity on SDR

In principle, due to its socially desirable nature, prosocial behavior as measured in economic games should be linked to SDR (Roussel, 2017). However, the degree to which behavior in economic games is affected by SDR is arguably influenced by the consequences and visibility of one's behavior, in the sense that individuals' responses may be more strongly shifted toward the socially desirable (i.e., prosocial) option when it does not actually affect their (monetary) outcomes and when it is more visible to others (Thielmann et al., 2016).

First, if behavior in games is merely hypothetical, SDR may affect responses more strongly than if behavior is incentivized and thus consequential. Specifically, if behavior is hypothetical, presenting oneself in a socially desirable way through prosocial behavior does not incur any (material) costs. By contrast, if behavior is real (i.e., incentivized), prosocial behavior is truly costly because it is, by definition, associated with forgoing personal gain to benefit others. That said, direct empirical comparisons (Amir et al., 2012; Thielmann et al., 2016) as well as meta-analytic evidence (Engel, 2011; Johnson & Mislin, 2011) suggest that the prevalence of prosocial behavior is largely comparable in incentivized versus hypothetical games, supporting that both these implementations of games provide valid measures of prosocial *behavior*. Likewise, the relation of personality traits to prosocial behavior is—for most traits—comparable in the absence versus presence of incentives (Thielmann et al., 2020). Nonetheless, it is conceivable that responses in hypothetical games are more strongly driven by SDR, whereas behavior-contingent incentives may decrease one's tendency to behave in a socially desirable manner (e.g., Baron, 2001). Thus, if the relation between SD scores and prosocial behavior is smaller when behavior in games is incentivized rather than hypothetical, this would support the style interpretation of SD scales. This is also in line with the findings by Müller and Moshagen (2019) summarized above.

Second, SDR should be affected by the level of anonymity of participants' responses, which may be affected by the setting of the study. In online settings, anonymity is increased

compared to lab or field experiments. Given that one underlying motivation of SDR is to avoid being negatively judged by others (Schaeffer, 2000), SDR can be expected to decrease with reduced opportunities to be identified as someone who has undesirable qualities, as is the case online. In line with this reasoning, the *candor hypothesis* (Buchanan, 2000) proposes that there is less SDR in surveys conducted online. Support for this claim comes from Gnambs and Kaspar (2015), who found that individuals were more inclined to report sensitive behaviors in computerized as compared to pen-and-pencil surveys.

Crucially, in lab settings, prosocial behavior may not only generally be less anonymous (and thus more driven by SDR) but providing behavior-contingent payment might further increase this tendency. This is because experimenters are potentially able to infer participants' selfishness based on their payoff unless double-blind payment is used (Cherry et al., 2002; Hoffman et al., 1999). For example, if a payoff is based on the outcome of a dictator game and a dictator receives the entire endowment as payoff, it is obvious that the participant kept all the money, giving nothing to the recipient. Thus, instead of decreasing SDR by using incentives, the effect might be reversed in lab settings. Conversely, a fully anonymous (e.g., online) and incentivized setting provides the most promising scenario in which behavior should—at least in theory—be unaffected by SDR. In the current meta-analysis, we therefore test the potential moderating effects of incentives and anonymity as well as their interaction on the relation between SD scores and prosocial behavior.

2 | METHOD

2.1 | Search for studies

The literature search involved three steps, as summarized in Figure S1 in the online supplement. First, we started with 23 documents reporting on data, including an SD scale and an economic game as identified in the data collection by Thielmann et al. (2020) for their meta-analysis on the link between personality and prosocial behavior in the six games focused on here. Second, in August and September 2019, we searched several databases, including Business Source Elite, PsycInfo, Web of Science, and Google Scholar using the search string ("Social Desirability" OR BIDR OR "Balanced Inventory of Desirable Responding" OR "Marlowe-Crowne" OR "Eysenck Lie Scale" OR "Self-Deception Questionnaire*" OR "Impression Management" OR "Self-Deceptive Enhancement") AND ("Economic Game*" OR "Trust Game" OR "Dictator Game" OR "Ultimatum Game" OR "Public Good* Game" OR "Common* Dilemma" OR "Commons Game" OR "Prisoner* Dilemma" OR "Voluntar* contribut* experiment*" OR "Voluntar* contribut* mechanism" OR "Social dilemma*" OR "Mixed-motive game*" OR "Resource dilemma*" OR "Common pool game" OR "Give-some dilemma" OR "Take-some dilemma" OR "Give-some game" OR "Take-some game" OR "Donat*" OR "Charity" OR "Charitable giving"). This yielded 1,234 results,[2] of which we excluded all non-English documents and duplicates. All remaining documents were screened with regard to whether they contained (1) an SD scale and (2) at least one of the six relevant economic games. If a document satisfied these criteria but did not contain the relevant (zero-order) correlations between the SD score and prosocial behavior, we contacted the corresponding author. In case the author did not reply within three weeks, we sent a reminder. For 11 documents, we received the requested data. In addition, one author shared another published dataset that was not yet included based on the initial search. For 13 documents, we did not receive any response from the corresponding author; therefore, these documents were excluded due to insufficient data.

Third, between mid-June and early July 2019, we sent out calls for (published or unpublished) data through several mailing lists. These included the Society of Judgment and Decision Making, the German Psychological Association, the European Association of Social Psychology, the European Association of Decision Making, the International Conference of Social Dilemmas, the Economic Science Association, and the European Association of Personality Psychology. Overall, five authors responded to our calls for data, which led to the inclusion of another published article that was not yet identified through the initial searches.

Overall, the meta-analysis included 32 documents reporting on data from 41 independent samples and a total of $N = 8980$ participants. Table 2 provides more detailed information on the involved samples, including sample sizes (N), SD scale, economic game(s), study characteristics (i.e., incentivization and setting) as well as effect sizes (i.e., Pearson's r). Further, all effects are summarized in the forest plot in Figure 1.

2.2 | Inclusion criteria

To be included in the meta-analysis, the reported studies had to fulfill the following criteria. First, all participants had to be adults, that is, at least 18 years of age. Second, studies had to contain a measure of SD in the form of a multi-item SD scale. Third, studies had to include a measure of prosocial behavior in one of the following six economic games: the dictator game, ultimatum game, trust game, prisoner's dilemma, public goods game, and commons dilemma. We included game settings in which other individuals were the interaction partners, as well as game settings in which charity organizations were the receivers.

TABLE 2 Studies included in the meta-analysis

Sample	Study number	*N*	Scale	Game	Incentive	Setting	*r*
Anderson et al. (2013)	NA	1261	Other	PGG	H	Offline	.02
Arnocky et al. (2017)	2	508	BIDR	DG	I	Offline	.05
Baran et al. (2010)	NA	477	MC	TGA, TGB, PGG, DG	I + H	Offline	.05
Baumert et al. (2014)	2	492	BIDR	DG, UGA, UGB	I	Online	.02
Bergsieker (2012)	2a	120	MC	PDG	H	Offline	.02
Bergsieker et al. (2018)	1	180	MC	PDG	H	Offline	.03
	2	159	MC	PDG	H	Offline	.11
	3	237	MC	PDG	H	Offline	.03
Corr et al. (2015)	NA	108	NA	PGG, TGA, TGB	H	Offline	.03
Dalbert and Umlauft (2009)	2	59	BIDR	DG	H	Offline	.41
DeCelles et al. (2012)	1	160	MC	DG	I	Online	−.09
Exline and Hill (2012)	1	195	MC	DG	I	Online	.01
	2	286	MC	DG	I	Online	.07
Fleming and Zizzo (2011)	NA	48	SDS-17	PGG	H	Offline	.42
Fleming and Zizzo (2015)	NA	64	SDS-17	DG	I	Online	.13
Franzen and Pointner (2013)	1a	138	Other	DG	I	Offline	.07
	1b	276	Other	DG	I	Offline	.08
Franzen and Mader (2019)	NA	365	MC	DG	I	Online	−.01
Gallucci and Perugini (2000)	NA	74	Other	DG	I	Offline	.07
Haring et al. (2013)	NA	55	Other	TGA	NA	Offline	.22
Ingram and Berger (1977)	NA	57	MC	PDG	H	Offline	.18
Konow and Earley (2008)	NA	48	MC	DG	I	Offline	−.09
Li et al. (2017)	NA	98	Others	DG, PDG	I	Offline	−.06
Maltese et al. (2016)	2	97	BIDR	TGA, TGB	I	Offline	.08
Margittai et al. (2015)	NA	78	SDS-17	DG	H	Offline	.12
Meleady and Seger (2017)	1	141	SDS-17	PDG	I	Online	−.16
	2	95	SDS-17	PDG	I	Online	−.20
	3	172	SDS-17	PDG	I	Online	.03
Müller and Moshagen (2019)	NA	460	BIDR	DG, PGG, CD	I + H	Online	.15
Näf and Schupp (2009)	NA	292	BIDR	TGB	I	Online	−.01
Nehrlich et al. (2019)	1	672	BIDR	DG, UGA, UGB	I	Offline	−.03
Perugini et al. (2003)	1	200	BIDR	UGA, UGB	H	Offline	.08
Raine and Uh (2019)	NA	297	Other	DG	I	NA	.02
Rêgo et al. (2017)	NA	19	MC	UGB	I	Online	.06
Rodrigues et al. (2015)	NA	40	MC	DG	H	Online	−.13
Smith (2016)	NA	117	BIDR	DG	I	Offline	−.04
Surbey and McNally (1997)	NA	140	Other	PDG	H	NA	.30
Surbey (2011)	NA	80	Other	PDG	H	NA	.21

(Continues)

TABLE 2 (Continued)

Sample	Study number	*N*	Scale	Game	Incentive	Setting	*r*
Uziel (2014)	1	72	BIDR	DG	I	Online	.03
	3	78	BIDR	DG	I	Offline	.22
Zhao et al. (2017)	3	304	SDS-17	DG	H	Online	.06
Zizzo and Fleming (2011)	NA	216	SDS-17	PGG	I	Online	−.03

Abbreviations: BIDR, balanced inventory of desirable responding; CD, commons dilemma; DG, dictator game; H, Hypothetical; I, incentivized; MC, Marlowe-Crowne; NA, no information on scale available; PDG, prisoner's dilemma game; PGG, public goods game; SDS-17, social desirability scale 17; TGA, trust game trustor; TGB, trust game trustee; UGA, ultimatum game proposer; UGB, ultimatum game responder.

2.3 | Coding of effect sizes

As an effect size, we used Pearson's correlation coefficient *r*. If a study contained several effect sizes relating SD to prosocial behavior (e.g., because IM and SDE were both measured with the BIDR but the SD score was not reported), we aggregated the respective effect sizes while, if available, taking into account the intercorrelations of the to-be-aggregated measures (e.g., the intercorrelation of IM and SDE; Hunter & Schmidt, 2004). This way, we ensured that we included each independent sample only once in our analysis.

For the overall meta-analytic aggregation of effects, we used random-effects psychometric meta-analysis with sample size weighting (Hunter & Schmidt, 2004). Furthermore, if Cronbach's alpha as a measure of internal consistency of an SD scale was available, we used alpha to correct for attenuation of effect sizes (Spearman, 1904).

2.4 | Coding of study characteristics

The first and second authors conducted the coding of studies. Table S2 in the online supplement provides an overview of all variables coded. In the following, we briefly elaborate on those variables that are relevant to the analyses presented here.

2.4.1 | Incentives

We coded whether a study measured prosocial behavior in a hypothetical or an incentivized game. Regarding the latter, incentives could be provided to all participants for all choices (full payment), to a random subset of participants (lottery payment), or a random subset of choices (randomized payment). Overall, nine studies used hypothetical games, whereas 29 studies used incentivized games. For one study, there was no information on incentives provided, and another two studies reported an aggregated effect of hypothetical and incentivized games, meaning that a separate assessment was not possible. Therefore, these studies were excluded from the moderation analyses.

2.4.2 | Setting

We coded whether a study was conducted online (i.e., on the internet) or offline (e.g., in the lab). In online studies, there is greater anonymity than in offline studies. Overall, 14 studies were conducted online, and 24 studies were conducted offline. For three studies, there was no information on the setting provided.

2.4.3 | SD scale

We coded the SD scale used to measure SDR. As summarized in Table 1, the most prevalent ones were the MC scale (and various short forms thereof; $k = 13$), the BIDR ($k = 11$), and the Social Desirability Scale-17 (SDS-17; Stöber, 2001; $k = 8$). Furthermore, we grouped the scales of nine studies into the category “others”, namely, the Self-Deception Questionnaire (Sackeim & Gur, 1979), the Eysenck Lie Scale (Eysenck & Eysenck, 1964), and several ad-hoc created scales that combined items from different SD scales. For one study, no information on the SD scale used was available.

2.4.4 | SD sub-dimensions

If the BIDR was used to measure SDR and separate scores for IM and SDE were reported, we coded separate effect sizes for both sub-dimensions of SD as well as for the total SD score (after aggregating effects for IM and SDE while taking their intercorrelations into account; see above). For six studies, information on the correlations between these sub-dimensions and prosocial behavior were available.

2.4.5 | Game type

We coded the type of game used to measure prosocial behavior. As mentioned above, we included six games (see Table S1 for details on these games). Most studies (i.e., $k = 23$) included the dictator game or social dilemmas[3] (i.e., $k = 17$). Of note, given that all games included measures of prosocial behavior—even though in different classes of social situations—we do not consider game type further in our analysis. This is because we focus on prosocial behavior as a kind of socially desirable behavior *in general*, rather than emphasizing the fine-grained differences between games.

3 | RESULTS

3.1 | Analytic procedure[4]

For all analyses, we used the *metafor* package (Viechtbauer, 2010) in R (R Core Team, 2018). As effect size, we calculated the meta-analytic, disattenuated, zero-order correlations between SD scores and prosocial behavior, using random-effects meta-analysis with sample size weights (Hunter & Schmidt, 2004). For those six samples for which it was possible, we additionally computed the meta-analytic correlations of the sub-dimensions—IM and SDE—with prosocial behavior. We used Cochran's Q to assess the extent of heterogeneity in effect sizes between studies, T^2 to assess the extent of between-study variance, and I^2 to assess the percentage of between-study variance that is attributable to true heterogeneity. To test for potential moderation effects of certain study characteristics (i.e., incentives, setting, incentives × setting, and SD scales), we used random-effects meta-regressions, predicting the disattenuated effect sizes by a dummy variable representing the respective moderator(s).[5]

To investigate the presence of publication bias, we used the rank correlation method (Begg & Mazumdar, 1994), Egger's regression test (Egger et al., 1997), and the trim-and-fill method (Duval & Tweedie, 2000a, 2000b). Importantly, there was no indication for publication bias on more than one of these indicators for any of the correlation analyses, supporting that the results were not systematically affected by publication bias. Thus, we only report the corresponding statistics in the online supplement (Table S3).

3.2 | Social desirability and prosocial behavior

Figure 1 shows the correlations between SD scores and prosocial behavior for all 40 samples included in the meta-analysis. Table 3 provides more detailed information, including exact p-values as well as Q, T^2, and I^2 statistics for the analysis of the overall SD score and the sub-dimensions IM and SDE. As is apparent, there was a weak, positive meta-analytic correlation between SD scores and prosocial behavior ($\hat{\rho} = .04$, $p = .006$, 95% CI [.01, .07]). Although the Q statistic did not reach significance ($Q = 50.78$, $p = .12$), there was evidence for some between-study variability in effect sizes due to true heterogeneity ($I^2 = 18.74\%$). This emphasizes the importance of investigating potential moderation by study characteristics.

For the two sub-dimensions of SD, IM and SDE, meta-analytic correlations showed the same pattern (Table 3), yielding weak positive effects for both IM ($\hat{\rho} = .07$, $p = .007$, 95% CI [.02, .12]) and SDE ($\hat{\rho} = .03$, $p = .307$, 95% CI [−.03, .09]). Overall, analyses thus showed positive links between SD and prosocial behavior that were, however, small at best. We report the correlations between IM and prosocial behavior and SDE and prosocial behavior for those studies that made a distinction possible in Figures S2 and S3, respectively.

3.3 | Moderation analyses

In addition to analyzing the overall correlation between SD and prosocial behavior, we also investigated whether this relation was moderated by three critical study characteristics, namely incentives, setting, and type of SD scale used. Results of the moderation analyses are summarized in Table 4.

3.3.1 | Incentives

To test whether the implementation of behavior-contingent incentives reduces the relation between SD scores and prosocial behavior, we predicted the disattenuated correlations by a dummy variable coding whether incentives were used (hypothetical = 0, incentivized = 1). Indeed, there was a significant moderation of the relation between SD and prosocial behavior by incentives, $b = -.10$, $SE = .02$, $p < .001$. Specifically, correlations between SD scores and prosocial behavior were stronger in hypothetical games ($\hat{\rho} = .07$, $p = .032$) than in incentivized games ($\hat{\rho} = .02$, $p = .159$). The findings were thus compatible with the style interpretation of SD scales: Once it becomes costly to present oneself in a positive way by behaving in a prosocial manner, the relation between SD scores and behavior decreases. Nonetheless, it should be noted that even when behavior was incentivized, the effect of SD scores was *not* negative, which was—strictly speaking—to-be-expected if SD scales would measure style rather than substance.

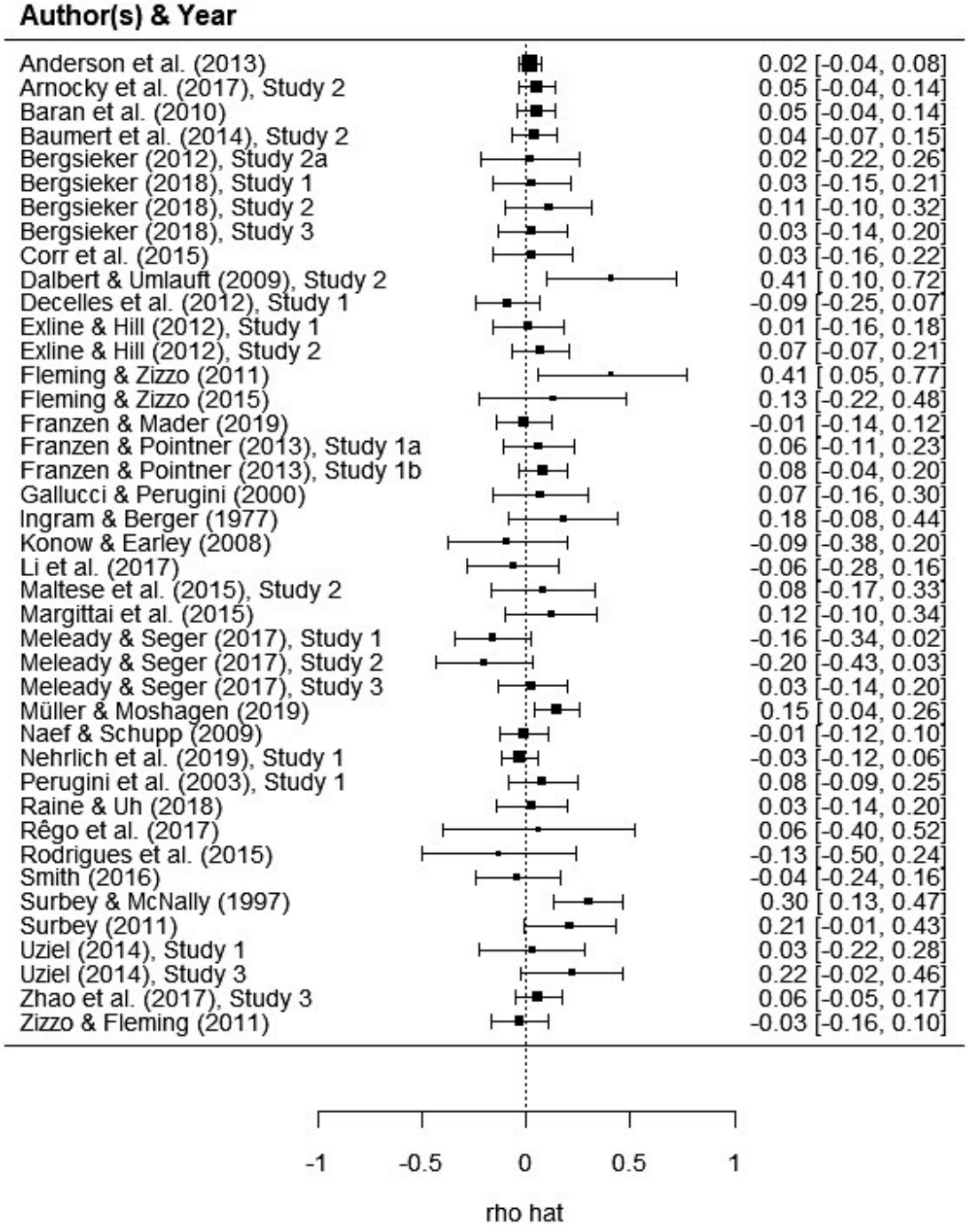

FIGURE 1 Forest plot of meta-analysis of relation between social desirability and prosocial behavior, with meta-analytic disattenuated correlations per study and 95% per study and 95% CL (in brackets)

TABLE 3 Meta-analysis of the relation between SD scores and prosocial behavior

Variable	*k*	*N*	$\hat{\rho}$ (*SE*)	95% CI	95% PI	*p*	*Q*	T^2	I^2	*r* (*SE*)
SD	41	8980	.04* (.01)	[.01, .07]	[−.04, .12]	.006	50.79	.001	18.85	.04* (.01)
IM	6	2429	.07* (.02)	[.03, .12]	[.02, .13]	.003	6.16	.000	2.49	.06* (.02)
SDE	6	2429	.03 (.03)	[−.03, .09]	[−.06, .12]	.307	8.28	.001	26.39	.02 (.02)

Abbreviations: CI, confidence interval; I^2, variation across samples due to true heterogeneity; IM, impression management; *k*, number of independent samples; *N*, total sample size; PI, prediction interval; *Q*, Cochran's *Q* statistic; *r*, mean (bare-bones) correlation; SD, social desirability; SDE, self-deceptive enhancement; *SE*, standard error; T^2, between-study variance; $\hat{\rho}$, mean true-score correlation corrected for unreliability.

*$p < .05$.

TABLE 4 Results of the moderator analyses on the relation between SD scores and prosocial behavior

Variable	*Q* (*df*)	*k*	*N*	$\hat{\rho}$ (*SE*)	95%CI
Incentivization	16.73* (1)				
Incentivized		29	5784	.02 (.02)	[−.01, .05]
Hypothetical		9	2620	.07* (.03)	[.01, .14]
Setting	0.71 (1)				
Online		14	3314	.02 (.02)	[−.02, .07]
Offline		24	5149	.04* (.02)	[.01, .07]
Incentive × Setting	5.07 (3)				
Incentivized × Online		12	2550	−.00 (.02)	[−.05, .04]
Incentivized × Offline		16	2937	.04* (.02)	[.00, .09]
Hypothetical × Online		2	764	.09* (.04)	[.01, .17]
Hypothetical × Offline		5	1636	.03 (.03)	[−.02, .08]
SD scale	2.69 (3)				
BIDR		11	3047	.05* (.03)	[.00, .10]
MC		13	2343	.03 (.02)	[−.02, .08]
SDS-17		8	1118	.01 (.05)	[−.08, .10]
Others		9	2872	.06* (.03)	[.00, .12]

Abbreviations: BIDR, balanced inventory of desirable responding; CI, confidence interval; *df*, degrees of freedom; *k*, number of effect sizes (in regression model); MC, Marlowe-Crowne scale; *N*, total sample size; *Q*, heterogeneity due to moderators; SDS-17, social desirability scale 17; $\hat{\rho}$, mean true-score correlation corrected for unreliability with standard error (*SE*).

*$p < .05$.

3.3.2 | Setting

We further tested whether the experimental setting (i.e., offline vs. online) —and resulting differences in anonymity— moderated the relation between SD and prosocial behavior. To this end, we predicted the disattenuated correlations by a dummy variable coding whether a sample was collected offline (=0) or online (=1).

The moderation analysis showed no significant effect of the experimental setting on the relation between SD scores and prosocial behavior, $b = -.02$, $SE = .03$, $p = .400$. Irrespective of whether data were collected online or offline, the correlations between SD scores and prosocial behavior were negligible (i.e., $\hat{\rho} = .02$; $p = .324$ and $\hat{\rho} = .04$, $p = .004$, respectively). This result is compatible with the substance interpretation of SD scales: Irrespective of whether prosocial behavior was more or less anonymous, SD scores were unrelated to prosocial behavior.

3.3.3 | Interaction between incentives and setting

In line with the argument that identifiability might reverse the effect of incentives on prosocial behavior and its relation to SD scores, we tested for a potential interaction between incentives and experimental setting. This allowed us to test whether the attenuating effect of incentives on the relation between SD scores and prosocial behavior may be reversed in offline settings in which the implementation of incentives can decrease anonymity.

In contrast to this reasoning, however, the moderation analysis showed no significant interaction between incentives and setting in predicting the correlations between SD scores and prosocial behavior, $b = -.09$, $SE = .06$, $p = .140$.

3.3.4 | SD scale

Because the SD scales used to measure SDR differ in terms of several aspects (see Table 1), we finally investigated the potentially moderating effect of the SD scale on the relation between SD scores and prosocial behavior. We used three dummy variables comparing the BIDR (as baseline) against MC, SDS-17, and other SD scales.

As summarized in Table 4, the meta-regression showed no moderation of the relation between SD scores and prosocial behavior by SD scale used, $Q(3) = 2.69$, $p = .442$. Specifically, two prevalent SD scales in our analysis, the MC-scale ($\hat{\rho} = .03$, $p = .259$), and SDS-17 ($\hat{\rho} = .01$, $p = .822$), yielded very small and non-significant effects, $b = -.03$, $SE = .04$, $p = .415$ for the MC-scale dummy and $b = -.05$, $SE = .04$, $p = .262$ for the SDS-17 dummy. The remaining "other" scales ($\hat{\rho} = .06$, $p = .042$) showed a small and significant positive correlation with prosocial behavior, yet again revealing no significant moderation, $b = .02$, $SE = .03$, $p = .590$. In turn, the BIDR showed significant

correlation ($\hat{\rho}$ = .05, p = .046) and a significant effect, b = .06, SE = .02, p = .013. Taken together, the moderation analysis thus provided no evidence for systematic differences between SD scores and prosocial behavior when using different SD scales.

4 | GENERAL DISCUSSION

Although there is a long history of using SD scales to measure SDR in psychology and beyond, it is—until this day—unclear what SD scales essentially measure. Whereas the debate about what SD scales measure has become more prominent in the literature in recent years, this discussion has mainly revolved around two different interpretations: a response bias (i.e., style) and substantive trait variance (i.e., substance). In addition, some evidence suggests that SD scales measure a blend of both substance and style (e.g., Müller & Moshagen, 2019). However, there is still a lack of large-scale evidence for any of these interpretations. Recognizing that method and theory are inextricably linked (van Maanen et al., 2007), we took the notion seriously that—in order to examine the validity of a measure thoroughly and, consequently, derive theoretical implications—a suitable methodology needs to be applied. To this end, we conducted a meta-analysis linking SD scale scores to prosocial behavior in economic games.

4.1 | Summary of results and theoretical implications

In studying the relation between SD scales and prosocial behavior in economic games, we overcome a limitation of prior research that has also been acknowledged by authors of SD scales (Paulhus & Vazire, 2007), namely, that (self-) report data may be prone to SDR, thus being insufficient as criterion for validation. In particular, we argued that a negative relation between SD scores and prosocial behavior favors the style interpretation of SD scales because it implies that individuals with high SD scores, irrespective of their absolute level of prosociality, systematically over-*report* having socially desirable attributes while failing to *behave* accordingly (i.e., in an equally prosocial manner). A positive correlation, by contrast, favors the subst—ance interpretation of SD scales because it implies that individuals with high SD scores not only report to have socially desirable attributes but indeed behave accordingly. Finally, a null correlation most likely suggests that SD scales do not clearly measure either substance or style but a blend of both.

Our results show that the correlation between SD scores and prosocial behavior is slightly positive but negligible in size ($\hat{\rho}$ = .04). In principle, a positive correlation would support the substance interpretation of SD scales; however, the (very) small effect size can hardly be considered as providing corresponding evidence. Although "an effect-size *r* of .05 indicates an effect that is very small for the explanation of single events but potentially consequential in the not-very long run" (Funder & Ozer, 2019, p. 156), we maintain that the effect observed in our meta-analysis should not be over-interpreted. Nonetheless—irrespective of how one interprets the size of the (positive) relation—if SD scales measured what they intend to measure, namely overly positive self-presentation, SD scores should have yielded a meaningful *negative* correlation with prosocial behavior. Thus, our findings generally question the validity of SD scales. Importantly, this conclusion cannot be attributed to the potential confounding of different sub-dimensions of SDR as implemented in some SD scales. When distinguishing between SDE and IM (as operationalized in the BIDR), the exact same picture emerged, namely close to zero correlations with prosocial behavior (i.e., $\hat{\rho}$ = .03 and $\hat{\rho}$ = .07, respectively). Thus, even when investigating SDR at a more fine-grained level, results still failed to support the style interpretation—and thus validity—of SD scales.

Second, we investigated the moderating role of different study characteristics on the relation between SD scores and prosocial behavior, namely (monetary) incentives and setting (i.e., offline vs. online), as well as their interaction. We argued that a weaker relation between SD scores and prosocial behavior when incentives were present and in offline (more anonymous) settings would support the style interpretation of SD scales since both are sought to reduce SDR in games (Baron, 2001; Camerer & Hogarth, 1999). First, incentives render SDR costly in the sense that prosocial behavior is associated with a reduced monetary outcome for the individual (Müller & Moshagen, 2019). In line with this reasoning, we indeed found a significant moderation by incentives, showing a positive correlation in hypothetical games ($\hat{\rho}$ = .07), but a lower, and close to zero, correlation in incentivized games ($\hat{\rho}$ = .02). In principle, this finding is compatible with the style interpretation of SD scales, suggesting that SD scores—at least to a certain extent—capture SDR. However, it is important to note that even in the incentivized games, the correlation between SD scores and prosocial behavior was still positive, whereas one would expect a negative correlation if SD scales measured style. Second, lab settings usually come with greater identifiability than online settings. Lab settings may thus increase SDR due to lower anonymity (e.g., Buchanan, 2000; Gnambs & Kaspar, 2015; Yaniv et al., 2020). However, there was no evidence for a moderation by online ($\hat{\rho}$ = .02) versus offline ($\hat{\rho}$ = .05) setting—and thus by differences in anonymity. Further, there was no evidence for an interaction between incentives and setting. These findings, again,

stay in conflict with the style interpretation of SD scales, which would have implied a stronger relation between SD scores and prosocial behavior once participants may fear that their responses are somehow identifiable, as may be the case in offline settings, particularly so if behavior is incentivized.

Lastly, we tested whether the usage of different SD scales affects the relation between SD scores and prosocial behavior. A significant moderation by SD scale would indicate that different scales cannot be considered equivalent, implying that different SD scales may measure different aspects of SDR. By implication, the non-equivalence of different SD scales may also indicate that these scales comply with either the style or the substance interpretation of SD scores to varying degrees. Overall, our analysis showed no significant moderation by the type of SD scale used. However, this lack of effect may also, to some extent, be attributable to the small sample sizes available for some scales. Nonetheless, these findings generally imply that irrespective of which SD scale one uses, no scale is suitable for measuring SDR reliably.

Taken together, our meta-analysis consistently shows that SD scales do not measure what they are supposed to, namely overly positive self-presentation. Whereas the observed close to zero correlation between SD scores and prosocial behavior does not provide clear support for either *one* interpretation, moderation analyses yielded evidence partly in favor of the style interpretation and partly in favor of the substance interpretation. As such, each of our individual findings individually might neither provide strong evidence for or against the style nor the substance interpretation, yet in conjunction, they further question the validity of SD scales as measures of SDR. Together, our findings this cast doubts on the usefulness of SD scales to uncover either *only* overly positive self-presentation in self-report data or *only* substantive traits—but rather a mixture of both. Future research is needed to clarify this issue ultimately.

4.2 | Practical implications

SD scales are widely used across fields in research and practice (e.g., personnel selection or clinical assessment of prisoners). Originally, scholars have used SD scales to measure individuals' tendency to present themselves overly positive. Our results suggest that this goal is not achieved by commonly used SD scales. Thus, it seems illegitimate to unequivocally interpret high SD scores as indicating "faking good". However, our results also contrast with more recent notions that SD scales mainly measure substantive traits, such as true virtue (de Vries et al., 2014; Zettler et al., 2015) or self-control (e.g., Uziel, 2010). Instead, SD scales may measure both—style and substance—meaning that they neither allow definite conclusions on individuals' tendency to present themselves overly positive nor on their trait characteristics. As such, our results substantiate that "a sufficient justification for the use of bias indicators in applied settings remains elusive" (McGrath et al., 2010, p. 450). Thus, echoing prior calls (e.g., Cunningham et al., 1994; de Vries et al., 2014, 2018), we recommend researchers and practitioners to refrain from using SD scales as measures of SDR but to instead rely on validated methods and measures to counteract SDR and assess individuals' (un)desirable trait characteristics.

Different approaches have been proposed to solve the issues resulting from SDR. One line of research aims at making participants believe that faking can be detected or, conversely, increases the level of anonymity to reduce self-presentational concerns. For example, the bogus pipeline technique (Jones & Sigall, 1971) makes participants believe that an anti-faking device (e.g., a lie detector) is used, thus involving deception by design—which is at odds with our field's ethics code (American Psychological Association, 2017). Alternatively, researchers interested in investigating sensitive issues that are prone to SDR, such as income data, drug abuse, or voting behavior (e.g., Brownback & Novotny, 2018; Epstein, 2006; Tourangeau & Yan, 2007), may rely on indirect questioning techniques such as the randomized response technique (RRT). RRT relies on a randomization device that instructs participants to either answer honestly or to respond in a certain way (e.g., in a socially undesirable way), regardless of what would have been the individual's honest answer. Consequently, the experimenter cannot distinguish between those who have chosen the socially undesirable option because it is their truthful response or because the instruction told them so and thereby enhances participants' anonymity. Meta-analytic evidence suggests that the RRT is an effective means to reduce SDR in surveys (Lensvelt-Mulders et al., 2005). However, the RRTs may also elicit suspicion and confusion among respondents (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018), suggesting that this method is also no all-purpose tool. Moreover, another drawback of the RRT is that it requires (very) large sample sizes and only allow drawing conclusions on the aggregate level.

To conclude, when the goal is to measure substantive trait variance related to socially desirable attributes, scholars should rely on validated personality scales rather than SD scales. Specifically, researchers interested in measuring virtuous characteristics may rely on the honesty-humility scale from the HEXACO Personality Inventory-Revised (Lee & Ashton, 2004). Researchers interested in measuring self-control—the other substantive trait that has been linked to SD scales (Uziel, 2010)—may, in turn, rely on well-validated self-control scales (e.g., Tangney et al., 2004). By this means, scholars can measure the substantive traits that have been

related to SDR without necessitating to interpret ambiguous SD scores.

4.3 | Limitations

By definition, meta-analytic research is restricted by the available data. In our meta-analysis, the number of independent samples was relatively small for some analyses or instances, such as for the distinction between SDE and IM ($k = 6$). Thus, statistical power to detect corresponding moderation effects may have been insufficient. Additionally, an even further differentiation beyond the IM/SDE distinction (see Paulhus, 2002) could have allowed to dig deeper into potential differences between sub-dimensions. Yet, data on such more fine-grained distinctions have not been incorporated in previous research. However, given that our analyses consistently questioned the style interpretation of SD scales, and because scholars seem to use SD scales more or less interchangeably, we are confident that our conclusions are at least somewhat generalizable. Likewise, the lack of relevant data puts certain limitations to the moderation analysis of the impact of anonymity. While SDR directed at others (e.g., IM) might be affected by anonymity, SDR directed at the self (e.g., SDE) is unconscious and therefore, likely not affected. Due to the very small number of datasets separating IM and SDE, we were not able to address these distinct mechanisms separately. Second, economic games are only one possible operationalization of prosocial behavior and there are other ways to measure prosocial behavior in experiments. For example, the social mindfulness paradigm specifically measures low-cost cooperation (van Doesum et al., 2013), and there are other donation or helping tasks that have been proposed to measure certain kinds of prosocial behavior in real-life contexts (see, e.g., Galizzi & Navarro-Martinez, 2019). Certainly, considering such paradigms in addition to economic games will provide more comprehensive insights into a person's prosociality. However, for the purpose of testing the validity of SD scales, the critical point is to have *a* measure that offers information on an individual's actual prosocial behavior, and this is exactly what economic games can accomplish and for what they have become a well-established research tool across fields (Thielmann et al., 2021; van Dijk & Dreu, 2021). Nonetheless, future research may test the generalizability of our findings to other measures of prosocial action. Finally, the incentivized settings (games) included in our analysis represented low-stakes situations—asking participants to allocate relatively small amounts of money—rather than high-stakes situations, such as job applications or clinical assessments of patients. While some studies have argued that low-stakes situations are related to lower levels of SDR (e.g., Lönnqvist et al., 2007; Mesmer-Magnus et al., 2006), Arthur et al. (2021) found in their review that SDR is prevalent in low- and high-stakes situations alike. In economic games in particular, the size of incentives seems to play a negligible role (e.g., Karagözoğlu & Urhan, 2017; Larney et al., 2019), as has also been shown by a study raising stakes to more than $1,600 (Johansson-Stenman et al., 2005). Even though this evidence supports the generalizability of our results to high-stakes situations, future research testing this assumption is needed. In any case, because SD scales are still widely used in psychology and beyond to control for SDR, it is crucial to know whether these scales are validly measuring style, irrespective of the stakes at hand.

4.4 | Conclusion

The results of this meta-analysis severely question the usefulness of SD scales as measures of "faking good" and overly positive self-presentation in surveys. At the same time, our results also present rather weak evidence for the alternative substance interpretation (i.e., considering SD scores as indicators of trait characteristics). These inconclusive results bear the risk that scholars and practitioners using SD scales cannot know for sure if they measure what they intend to measure (i.e., substance or style), or if their results are heavily confounded by the respective other, substantially different, option. We conclude that researchers and practitioners should refrain from using SD scales and, whenever possible, rely on alternative means to obtain survey responses that are not diluted by SDR.

ACKNOWLEDGMENTS

The authors thank Reinout de Vries for his valuable comments on the manuscript, the participants of the DGPs Writing Workshop 2020 for their insightful feedback, as well as all authors who contributed data or additional information on their studies.

CONFLICT OF INTEREST

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

ETHICS STATEMENT

At the authors' institutions, meta-analyses do not require ethics approval given that they only include data from other research projects for which ethics approval should (if necessary) have been obtained by the individual authors.

AUTHOR CONTRIBUTIONS

Conceptualization/Research Idea: Isabel Thielmann, Fabiola H. Gerpott, and Lukas Lanz; Data Collection & Analysis: Lukas Lanz and Isabel Thielmann; Manuscript Writing &

Editing: Lukas Lanz, Isabel Thielmann, and Fabiola H. Gerpott.

ORCID

Lukas Lanz https://orcid.org/0000-0001-5515-4873
Isabel Thielmann https://orcid.org/0000-0002-9071-5709
Fabiola H. Gerpott https://orcid.org/0000-0002-2585-3427

ENDNOTES

[1] Evidently, the only exception would be a person who behaves maximally prosocial and thus cannot present themselves even more positively when responding to the items of an SD scale (i.e., a ceiling effect).

[2] In an additional search in Spring 2021 extending the search scope of the SD scales by "MMPI" OR "Basic Personality Inventory" OR "Personality Research Form" OR MPQ OR "NEO Personality Inventory" OR "NEO PI" OR "NEO PI-R" OR PDQ-4 OR DS36 OR "Unlikely Virtues Scale", we found 292 results of which four turned out to be possibly relevant. Because none of these articles reported the effect sizes directly, we contacted the author of which two provided data which we included in the analyses.

[3] We grouped the prisoner's dilemma game, the public goods game, and the commons dilemma under the label of social dilemmas, given their structural similarity.

[4] The study was not pre-registered. All data and analysis scripts used in the meta-analysis are provided online on the Open Science Framework (https://osf.io/p9myc/).

[5] To rule out potential confounding effects of other study variables, we repeated all meta-regressions when including additional study characteristics (e.g., game type, repetition of interaction) as predictors in the regression model. All results remained essentially the same. Thus, we report results from these multivariate regression models in the online supplement only (Table S4).

REFERENCES

References marked by an asterisk (*) are included in the meta-analysis.

American Psychological Association. (2017). Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code/ethics-code-2017.pdf

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the internet: The effect of $1 stakes. *PLoS ONE*, *7*(2), e31461. https://doi.org/10.1371/journal.pone.0031461

*Anderson, J., Burks, S. V., Carpenter, J., Götte, L., Maurer, K., Nosenzo, D., Potter, R., Rocha, K., & Rustichini, A. (2013). Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools: Evidence from one college student and two adult samples. *Experimental Economics*, *16*(2), 170–189. https://doi.org/10.1007/s10683-012-9327-7

*Arnocky, S., Piché, T., Albert, G., Ouellette, D., & Barclay, P. (2017). Altruism predicts mating success in humans. *British Journal of Psychology*, *108*(2), 416–435. https://doi.org/10.1111/bjop.12208

Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. *Advances in Experimental Social Psychology*, *4*, 1–34. https://doi.org/10.1016/S0065-2601(08)60075-1

Arthur, W., Hagen, E., & George, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*(1), 105–137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Balliet, D., & van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, *139*(5), 1090–1112. https://doi.org/10.1037/a0030939

*Baran, N., Sapienza, P. & Zingales, L. (2010). Can we infer social preferences from the lab? Evidence from the trust game. National Bureau of Economic Research (NBER). https://doi.org/10.3386/w15654

Baron, J. (2001). Purposes and methods. *Behavioral and Brain Sciences*, *24*(3), 403. https://doi.org/10.1017/S0140525X01224145

*Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic games: A performance-based assessment of fairness and altruism. *European Journal of Psychological Assessment*, *30*(3), 178–192. https://doi.org/10.1027/1015-5759/a000183

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088. https://doi.org/10.2307/2533446

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The nature of faking: A homogeneous and predictable construct? *Psychological Assessment*, *31*(4), 532–544. https://doi.org/10.1037/pas0000619

Bensch, D., Paulhus, D. L., Stankov, L., & Ziegler, M. (2019). Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment*, *26*(3), 351–363. https://doi.org/10.1177/1073191117700268

*Bergsieker, H. B. (2012). *Building, betraying and buffering trust in interracial and same-race friendships* [Unpublished doctoral dissertation]. Princeton University.

*Bergsieker, H. B., Mu, F., & Cyr, E. N. (2018). *Inducing trust in interracial relationships: rewards of risky interdependence* [Manuscript in preparation].

Brownback, A., & Novotny, A. (2018). Social desirability bias and polling errors in the 2016 presidential election. *Journal of Behavioral and Experimental Economics*, *74*, 38–56. https://doi.org/10.1016/j.socec.2018.03.001

Buchanan, T. (2000). Potential of the internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 121–140). Academic Press. https://doi.org/10.1016/B978-012099980-4/50006-X

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A. M., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. University of Minnesota Press.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk & Uncertainty*, *19*, 7–42. https://doi.org/10.1023/A:1007850605129

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, *92*(4), 1218–1221. https://doi.org/10.1257/00028280260344740

Connelly, B. S., & Chang, L. (2016). A meta-analytic multitrait multirater separation of substance and style in social desirability scales. *Journal of Personality*, *84*(3), 319–334. https://doi.org/10.1111/jopy.12161

*Corr, P. J., Hargreaves Heap, S. P., Seger, C. R., & Tsutsui, K. (2015). An experiment on individual 'parochial altruism' revealing no connection between individual 'altruism' and individual 'parochialism'. *Frontiers in Psychology*, *6*, 1–8. https://doi.org/10.3389/fpsyg.2015.01261

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

Cunningham, M. R., Wong, D. T., & Barbee, A. P. (1994). Self-presentation dynamics on overt integrity tests: Experimental studies of the Reid report. *Journal of Applied Psychology*, *79*(5), 643–658. https://doi.org/10.1037/0021-9010.79.5.643

*Dalbert, C., & Umlauft, S. (2009). The role of the justice motive in economic decision making. *Journal of Economic Psychology*, *30*(2), 172–180. https://doi.org/10.1016/j.joep.2008.07.006

de Vries, R. E., Hilbig, B. E., Zettler, I., Dunlop, P. D., Holtrop, D., Lee, K., & Ashton, M. C. (2018). Honest people tend to use less-not more-profanity: Comment on Feldman et al.'s (2017) Study 1. *Social Psychological and Personality Science*, *9*(5), 516–520. https://doi.org/10.1177/1948550617714586

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. https://doi.org/10.1177/1073191113504619

*DeCelles, K. A., DeRue, D. S., Margolis, J. D., & Ceranic, T. L. (2012). Does power corrupt or enable? When and why power facilitates self-interested behavior. *Journal of Applied Psychology*, *97*(3), 681–689. https://doi.org/10.1037/a0026811

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89–98. https://doi.org/10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. The Dryden Press.

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

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583–610. https://doi.org/10.1007/s10683-011-9283-7

Epstein, W. M. (2006). Response bias in opinion polls and American social welfare. *The Social Science Journal*, *43*(1), 99–110. https://doi.org/10.1016/j.soscij.2005.12.010

*Exline, J. J., & Hill, P. C. (2012). Humility: A consistent and robust predictor of generosity. *The Journal of Positive Psychology*, *7*(3), 208–218. https://doi.org/10.1080/17439760.2012.671348

Eysenck, H. J., & Eysenck, S. B. G. (1964). *Manual of the Eysenck personality inventory*. University of London Press.

Feldman, G. (2019). What is honesty? Laypersons interpret high lie scale scores as reflecting intentional dishonesty: Rejoinder to de Vries et al.'s (2017) Comment on Feldman et al. (2017). *Social Psychological and Personality Science*, *10*(2), 220–226. https://doi.org/10.1177/1948550617737141

Feldman, G., Lian, H., Kosinski, M., & Stillwell, D. (2017). Frankly, we do give a damn: The relationship between profanity and honesty. *Social Psychological and Personality Science*, *8*(7), 816–826. https://doi.org/10.1177/1948550616681055

*Fleming, P., & Zizzo, D. J. (2011). Social desirability, approval and public good contribution. *Personality and Individual Differences*, *51*(3), 258–262. https://doi.org/10.1016/j.paid.2010.05.028

*Fleming, P., & Zizzo, D. J. (2015). A simple stress test of experimenter demand effects. *Theory and Decision*, *78*(2), 219–231. https://doi.org/10.1007/s11238-014-9419-2

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347–369. https://doi.org/10.1006/game.1994.1021

*Franzen, A., & Mader, S. (2019). Do phantom questions measure social desirability? *Methods, Data, Analyses*, *13*(1), 37–57. https://doi.org/10.12758/MDA.2019.01

*Franzen, A., & Pointner, S. (2013). The external validity of giving in the dictator game. *Experimental Economics*, *16*(2), 155–169. https://doi.org/10.1007/s10683-012-9337-5

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Galizzi, M. M., & Navarro-Martinez, D. (2019). On the external validity of social preference games: A systematic lab-field study. *Management Science*, *65*(3), 976–1002. https://doi.org/10.1287/mnsc.2017.2908

*Gallucci, M., & Perugini, M. (2000). An experimental test of a game-theoretical model of reciprocity. *Journal of Behavioral Decision Making*, *13*(4), 367–389. https://doi.org/10.1002/1099-0771(200010/12)13:4<367:AID-BDM357>3.0.CO;2-9

Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, *47*(4), 1237–1259. https://doi.org/10.3758/s13428-014-0533-4

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment*, *11*(4), 340–344. https://doi.org/10.1111/j.0965-075X.2003.00256.x

*Haring, K. S., Matsumoto, Y., & Watanabe, K. (2013). How do people perceive and trust a lifelike robot. *Proceedings of the World Congress on Engineering and Computer Science*, *1*, 425–430. http://www.iaeng.org/publication/WCECS2013/WCECS2013_pp425-430.pdf

Harris, S., Mussen, P., & Rutherford, E. (1976). Some cognitive, behavioral and personality correlates of maturity of moral judgment. *The Journal of Genetic Psychology*, *128*(1), 123–135. https://doi.org/10.1080/00221325.1976.10533980

Hathaway, S. R., & McKinley, J. C. (1951). *Minnesota multiphasic personality inventory manual, revised*. Psychological Corporation.

Hebert, J. R., Ma, Y., Clemow, L., Ockene, I. S., Saperia, G., Stanek, E. J., Merriam, P. A., & Ockene, J. K. (1997). Gender differences in social desirability and social approval bias in dietary self-report. *American Journal of Epidemiology*, *146*(12), 1046–1055. https://doi.org/10.1093/oxfordjournals.aje.a009233

Hoffman, E., McCabe, K., & Smith, V. L. (1999). Social distance and other-regarding behavior in dictator games: Reply. *American Economic Review*, *89*(1), 340–341. https://doi.org/10.1257/aer.89.1.340

Höglinger, M., & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: False positives undermine the crosswise-model RRT. *Political Analysis*, *25*(1), 131–137. https://doi.org/10.1017/pan.2016.5

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PLoS ONE*, *13*(8), e0201770. https://doi.org/10.1371/journal.pone.0201770

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality & Social Psychology Bulletin*, *30*(2), 161–172. https://doi.org/10.1177/0146167203259930

Holtrop, D., Hughes, A. W., Dunlop, P. D., Chan, J., & Steedman, G. (2020). Do social desirability scales measure dishonesty? *European Journal of Psychological Assessment*, 1–9. https://doi.org/10.1027/1015-5759/a000607

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). SAGE Publications, Inc. https://doi.org/10.4135/9781412985031

*Ingram, B. L., & Berger, S. E. (1977). Sex-role orientation, defensiveness, and competitiveness in women. *Journal of Conflict Resolution*, *21*(3), 501–518. https://doi.org/10.1177/002200277702100307

Johansson-Stenman, O., Mahmud, M., & Martinsson, P. (2005). Does stake size matter in trust games? *Economics Letters*, *88*(3), 365–369. https://doi.org/10.1016/j.econlet.2005.03.007

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865–889. https://doi.org/10.1016/j.joep.2011.05.007

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*(5), 349–364. https://doi.org/10.1037/h0031617

Karagözoğlu, E., & Urhan, Ü. B. (2017). The effect of stake size in experimental bargaining and distribution games: A survey. *Group Decision and Negotiation*, *26*(2), 285–325. https://doi.org/10.1007/s10726-016-9490-x

*Konow, J., & Earley, J. (2008). The Hedonistic Paradox: Is homo economicus happier? *Journal of Public Economics*, *92*(1–2), 1–33. https://doi.org/10.1016/j.jpubeco.2007.04.006

Lambert, C. E., Arbuckle, S. A., & Holden, R. R. (2016). The Marlowe-Crowne social desirability scale outperforms the BIDR impression management scale for identifying fakers. *Journal of Research in Personality*, *61*, 80–86. https://doi.org/10.1016/j.jrp.2016.02.004

Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *151*, 61–72. https://doi.org/10.1016/j.obhdp.2019.01.002

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

Lee, Z., & Sargeant, A. (2011). Dealing with social desirability bias: An application to charitable giving. *European Journal of Marketing*, *45*(5), 703–719. https://doi.org/10.1108/03090561111119994

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology*, *98*(4), 668–682. https://doi.org/10.1037/a0018771

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research. *Sociological Methods & Research*, *33*(3), 319–348. https://doi.org/10.1177/0049124104268664

*Li, X., Zhu, P., Yu, Y., Zhang, J., & Zhang, Z. (2017). The effect of reciprocity disposition on giving and repaying reciprocity behavior. *Personality and Individual Differences*, *109*, 201–206. https://doi.org/10.1016/j.paid.2017.01.007

Liu, X., & Liu, X. (2021). How social desirability influences the association between extraversion and reactive aggression: A suppression effect study. *Personality and Individual Differences*, *172*, 110585. https://doi.org/10.1016/j.paid.2020.110585

Lönnqvist, J.-E., Paunonen, S., Tuulio-Henriksson, A., Lönnqvist, J., & Verkasalo, M. (2007). Substance and style in socially desirable responding. *Journal of Personality*, *75*(2), 291–322. https://doi.org/10.1111/j.1467-6494.2006.00440.x

*Maltese, S., Baumert, A., Schmitt, M. J., & MacLeod, C. (2016). How victim sensitivity leads to uncooperative behavior via expectancies of injustice. *Frontiers in Psychology*, *6*, 1–11. https://doi.org/10.3389/fpsyg.2015.02059

*Margittai, Z., Strombach, T., vanWingerden, M., Joëls, M., Schwabe, L., & Kalenscher, T. (2015). A friend in need: Time-dependent effects of stress on social discounting in men. *Hormones and Behavior*, *73*, 75–82. https://doi.org/10.1016/j.yhbeh.2015.05.019

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

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*(3), 450–470. https://doi.org/10.1037/a0019216

*Meleady, R., & Seger, C. R. (2017). Imagined contact encourages prosocial behavior towards outgroup members. *Group Processes & Intergroup Relations*, *20*(4), 447–464. https://doi.org/10.1177/1368430215612225

Mesmer-Magnus, J., Viswesvaran, C., Deshpande, S., & Joseph, J. (2006). Social desirability: The role of over-claiming, self-esteem, and emotional intelligence. *Psychology Science*, *48*(3), 336–356.

*Müller, S., & Moshagen, M. (2019). True virtue, self-presentation, or both? A behavioral test of impression management and overclaiming. *Psychological Assessment*, *31*(2), 181–191. https://doi.org/10.1037/pas0000657

Murnighan, J. K., & Wang, L. (2016). The social world as an experimental game. *Organizational Behavior and Human Decision Processes*, *136*, 80–94. https://doi.org/10.1016/j.obhdp.2016.02.003

*Näf, M., & Schupp, J. (2009). *Can we trust the trust game? A comprehensive examination* [Discussion Paper Series 2009-5]. Royal Holloway University of London. https://intranet.royalholloway.ac.uk/economics/documents/pdf/discussionpapers/2009/dpe0905.pdf

*Nehrlich, A. D., Gebauer, J. E., Sedikides, C., & Schoel, C. (2019). Agentic narcissism, communal narcissism, and prosociality. *Journal of Personality and Social Psychology*, *117*(1), 142–165. https://doi.org/10.1037/pspp0000190

Ng, W. K., Shaban, R. Z., & van de Mortel, T. (2020). Hand hygiene beliefs and behaviours about alcohol-based hand rub use: Questionnaire development, piloting and validation. *Infection, Disease & Health*, *25*(1), 43–49. https://doi.org/10.1016/j.idh.2019.10.001

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660–679. https://doi.org/10.1037/0021-9010.81.6.660

Parks-Leduc, L., Feldman, G., & Bardi, A. (2015). Personality traits and personal values: A meta-analysis. *Personality and Social Psychology*, *19*(1), 3–29. https://doi.org/10.1177/1088868314538548

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/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 social psychological attitudes: Vol. 1. Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

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). Erlbaum. https://doi.org/10.4324/9781410607454-10

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*(6), 1025–1060. https://doi.org/10.1111/1467-6494.00041

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). The Guilford Press.

Perinelli, E., & Gremigni, P. (2016). Use of social desirability scales in clinical psychology: A systematic review. *Journal of Clinical Psychology*, *72*(6), 534–551. https://doi.org/10.1002/jclp.22284

*Perugini, M., Gallucci, M., Presaghi, F., & Ercolani, A. P. (2003). The personal norm of reciprocity. *European Journal of Personality*, *17*(4), 251–283. https://doi.org/10.1002/per.474

Platow, M. J. (1994). An evaluation of the social desirability of prosocial self-other allocation choices. *The Journal of Social Psychology*, *134*(1), 61–68. https://doi.org/10.1080/00224545.1994.9710884

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

*Raine, A., & Uh, S. (2019). The selfishness questionnaire: Egocentric, adaptive, and pathological forms of selfishness. *Journal of Personality Assessment*, *101*(5), 503–514. https://doi.org/10.1080/00223891.2018.1455692

*Rêgo, G. G., Campanhã, C., do Egito, J. H. T., & Boggio, P. S. (2017). Taking it easy when playing ultimatum game with a Down syndrome proposer: Effects on behavior and medial frontal negativity. *Social Neuroscience*, *12*(5), 530–540. https://doi.org/10.1080/17470919.2016.1195772

*Rodrigues, J., Ulrich, N., & Hewig, J. (2015). A neural signature of fairness in altruism: A game of theta? *Social Neuroscience*, *10*(2), 192–205. https://doi.org/10.1080/17470919.2014.977401

Roussel, S. (2017). Prosocial behaviors. In A. Marciano, & G. B. Ramello (Eds.), *Encyclopedia of law and economics* (pp. 1–4). Springer. https://doi.org/10.1007/978-1-4614-7883-6_710-1

Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology*, *47*(1), 213–215. https://doi.org/10.1037/0022-006X.47.1.213

Schaeffer, N. C. (2000). Asking questions about threatening topics: A selective overview. In A. A. Stone (Ed.), *The science of self-report: Implications for research and practice* (pp. 105–121). Lawrence Erlbaum.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-serving justifications. *Current Directions in Psychological Science*, *24*(2), 125–130. https://doi.org/10.1177/0963721414553264

Shalvi, S., Handgraaf, M. J. J., & de Dreu, C. K. (2011). Ethical manoeuvring: Why people avoid both major and minor Lies. *British Journal of Management*, *22*, S16–S27. https://doi.org/10.1111/j.1467-8551.2010.00709.x

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review*, *12*(3), 248–279. https://doi.org/10.1177/1088868308319226

*Smith, J. (2016). *Three essays on the origins and consequences of public service motives* [Doctoral Dissertation, Syracuse University]. ProQuest Dissertations Publishing.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin*, *134*(1), 138–161. https://doi.org/10.1037/0033-2909.134.1.138

Stöber, J. (2001). The social desirability scale-17 (SDS-17). *European Journal of Psychological Assessment*, *17*(3), 222–232. https://doi.org/10.1027//1015-5759.17.3.222

*Surbey, M. K. (2011). Adaptive significance of low levels of self-deception and cooperation in depression. *Evolution and Human Behavior*, *32*(1), 29–40. https://doi.org/10.1016/j.evolhumbehav.2010.08.009

*Surbey, M. K., & McNally, J. J. (1997). Self-deception as a mediator of cooperation and defection in varying social contexts described in the iterated prisoner's dilemma. *Evolution and Human Behavior*, *18*(6), 417–435. https://doi.org/10.1016/S1090-5138(97)00090-1

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Thielmann, I., Böhm, R., & Hilbig, B. E. (2015). Different games for different motives: Comment on Haesevoets, Folmer, and Van Hiel (2015). *European Journal of Personality*, *29*(4), 506–508. https://doi.org/10.1002/per.2007

Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021). Economic games: An introduction and guide for research. *Collabra: Psychology*, *7*(1), Article 19004. https://doi.org/10.1525/collabra.19004

Thielmann, I., Heck, D. W., & Hilbig, B. E. (2016). Anonymity and incentives: An investigation of techniques to reduce socially desirable responding in the trust game. *Judgment and Decision Making*, *11*(5), 527–536. http://journal.sjdm.org/16/16613/jdm16613.pdf

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30–90. https://doi.org/10.1037/bul0000217

Thielmann, I., Zimmermann, J., Leising, D., Hilbig, B. E., & Back, M. (2017). Seeing is knowing: On the predictive accuracy of self- and informant reports for prosocial and moral behaviours. *European Journal of Personality*, *31*(4), 404–418. https://doi.org/10.1002/per.2112

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

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

*Uziel, L. (2014). Impression management ("lie") scales are associated with interpersonally oriented self-control, not other-deception. *Journal of Personality*, *82*(3), 200–212. https://doi.org/10.1111/jopy.12045

van Dijk, E., & de Dreu, C. K. W. (2021). Experimental games and social decision making. *Annual Review of Psychology*, *72*, 415–438. https://doi.org/10.1146/annurev-psych-081420-110718

van Doesum, N. J., van Lange, D. A. W., & van Lange, P. A. M. (2013). Social mindfulness: Skill and will to navigate the social world.

Journal of Personality and Social Psychology, *105*(1), 86–103. https://doi.org/10.1037/a0032540

van Maanen, J., Sørensen, J. B., & Mitchell, T. R. (2007). The interplay between theory and method. *Academy of Management Review*, *32*(4), 1145–1154. https://doi.org/10.5465/amr.2007.26586080

Vecina, M. L., Chacón, F., & Pérez-Viejo, J. M. (2016). Moral absolutism, self-deception, and moral self-concept in men who commit intimate partner violence: A comparative study with an opposite sample. *Violence Against Women*, *22*(1), 3–16. https://doi.org/10.1177/1077801215597791

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Ward, S. J., & King, L. A. (2018). Religion and moral self-image: The contributions of prosocial behavior, socially desirable responding, and personality. *Personality and Individual Differences*, *131*, 222–231. https://doi.org/10.1016/j.paid.2018.04.028

Wiggins, J. S. (1964). Convergences among stylistic response measures from objective personality tests. *Educational and Psychological Measurement*, *24*(3), 551–562. https://doi.org/10.1177/001316446402400310

Williams, E. A., Pillai, R., Lowe, K. B., Jung, D., & Herst, D. (2009). Crisis, charisma, values, and voting behavior in the 2004 presidential election. *The Leadership Quarterly*, *20*(2), 70–86. https://doi.org/10.1016/j.leaqua.2009.01.002

Yaniv, G., Tobol, Y., & Siniver, E. (2020). Self-portrayed honesty and behavioral dishonesty. *Ethics & Behavior*, *30*(8), 617–627. https://doi.org/10.1080/10508422.2019.1678162

Zettler, I., Hilbig, B. E., Moshagen, M., & de Vries, R. E. (2015). Dishonest responding or true virtue? A behavioral test of impression management. *Personality and Individual Differences*, *81*, 107–111. https://doi.org/10.1016/j.paid.2014.10.007

*Zhao, K., Ferguson, E., & Smillie, L. D. (2017). Individual differences in good manners rather than compassion predict fair allocations of wealth in the dictator game. *Journal of Personality*, *85*(2), 244–256. https://doi.org/10.1111/jopy.12237

*Zizzo, D. J., & Fleming, P. (2011). Can experimental measures of sensitivity to social pressure predict public good contribution? *Economics Letters*, *111*(3), 239–242. https://doi.org/10.1016/j.econlet.2011.02.021

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

How to cite this article: Lanz, L., Thielmann, I., & Gerpott, F. H. (2021). Are social desirability scales desirable? A meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *Journal of Personality*, 00, 1–19. https://doi.org/10.1111/jopy.12662