



LUNDS
UNIVERSITET

Department of Psychology

**Thou Shalt Not Bear False Witness: An Exploratory Pilot
Study of the Relationships Between Social Desirability,
Honesty, and Self-Reported Past Behaviour in the Context
of Big Five Personality Research**

Ida Zdilar

Master's Thesis

2026

Supervisors: Fredrik Björklund, Martin Bäckström

Abstract

Contemporary efforts made toward creating a personality inventory free of the evaluative method variance, partially explained by social desirability, used the evaluative neutralisation method (Bäckström & Björklund, 2013; 2020; Bäckström et al., 2009; 2014; 2023). However, recent developments in the interpretation of the social desirability bias seem to suggest that there is perhaps reason to consider keeping said variance, as there appears to be some truth behind the alleged bias — SDR seems to both correlate positively with HEXACO's Honesty-Humility and improve criterion validity of personality inventories. The present exploratory pilot study ventured to construct a trait honesty measure and examine if it would disprove of the view of SDR as method variance and instead explain the substance of SDR by association, as well as if said measure would moderate the prediction of socially desirable behaviour based on SDR. An online sample (N = 299) was presented with a battery of tests needed to answer these research questions. I confirmed the existence of the evaluative factor, explained partially by SDR, within the traditional IPIP personality inventory. I also managed to predict socially desirable behaviour based on socially desirable responding. However, failure to find the moderation effect of honesty on the relationship between socially desirable responding and socially desirable behaviour steered future research in other directions, based on previous research — namely towards morality.

Key words: personality psychology, evaluative neutralisation, evaluative factor, honesty, social desirability, socially desirable behaviour

Acknowledgements

I would like to take this opportunity offer my gratitude first and foremost to my two wonderful supervisors, Fredrik Björklund and Martin Bäckström, who chose me instead of the other way around, and supported me throughout writing this thesis by providing me with both confidence in myself and unending psychometric support and patience. I would also like to acknowledge Roger Johansson for his vergilian presence at the end of this journey and for all the help and advice he has given me. Finally, I would like to thank my two sisters; Mika, newly Swedish citizen, who has made my move to and life in Sweden impossibly smoother than it would have been without her, and Buba, for always being in my corner even if I, myself, was not in it.

Introduction

The basis of psychological research lies, among other things, in the assumption that the instruments used indeed measure what they were intended to — that is, in construct validity. However, as is often the case with matters concerning human nature, and hence social science as a whole, absolute clarity of the measurement is rarely obtained. Regarding the self-report method specifically, the main obstacle of the instruments reflecting purely the intended construct is the tendency of participants to include a degree of inauthenticity in their responses, one which is commonly referred to as *self-presentation* (Paulhus & Trapnell, 2008). This propensity to presenting oneself in a better-than-is-realistic light was reported in many areas: from simple survey sampling and clinical contexts, to charitable giving and personnel selection (Gittelman et al., 2015; Lee & Sargeant, 2011; McGee Ng et al., 2015; Ones & Viswesvaran, 1998). At first glance, it may seem as though this phenomenon manifests only in certain situations. In actuality, there is reason to believe that it might represent trait-like individual differences, perhaps amplified, but certainly not exclusive to, instances in which one is overtly motivated to leave a good impression (Paulhus & Trapnell, 2008). Building onto this logic, researchers began thinking in terms of *socially desirable responding* (SDR), or *social desirability bias*, as a semi-stable idiosyncratic response style resulting in dissimulated scores on self-report questionnaires (Paulhus & Trapnell, 2008).

One of the areas in which SDR measurements are the most applicable is personality research. While some researchers claim social desirability does not affect neither factor structure of the Big Five nor the general factor of personality, maintaining the conviction that construct and criterion validity of personality measures are not in question (Ellingson et al., 2001; Marshall et al., 2005; Pelt et al., 2019), a substantial number of studies demonstrated the issue of social desirability bias influencing scores on multiple Big Five questionnaires. Those with a profile low on Neuroticism and high on the other four factors in a NEO-Five Factor Inventory (NEO-FFI) showed the highest correlation with social desirability measured by the Balanced Inventory of Desirable Responding (BIDR) (Roth & Herzberg, 2007), but relationships were also found when looking individually at specific factors and social desirability, namely positive correlations with agreeableness, extraversion, conscientiousness, and to a lesser extent openness to experience, as well as negative correlations with neuroticism (Bäckström & Björklund, 2013; De Jonge & Slaets, 2005; Dodaj, 2012; Dunlop et al., 2012; Havan & Kohut, 2019; Konstabel et al., 2006; McKelvie, 2004). These correlations have been shown to impact the factor structure of the Big Five model by

collapsing the assumption of orthogonality of the factors unless social desirability is accounted for statistically, as items most affected by social desirability tend to correlate more strongly regardless of their content (Navarro-González et al., 2016). In a study examining the factorial structures of the IPIP and NEO-FFI, a model that included a common method factor orthogonal to the five factors of personality, along with two separate item wording factors with correlations restricted to zero, emerged as superior in CFA analyses (Biderman et al., 2011). According to the authors, this factor seemed to represent individual differences in reporting on desirable content of the items; excluding the influence of these factors also lowered personality correlations with self-esteem and both positive and negative affectivity. Biderman and his associates are not the only ones to connect intercorrelations within the Big Five to social desirability. The first principal component extracted from a personality inventory was found to have had moderate to high correlations to both impression management and self-deception facets of social desirability, and the inventory that included it showed higher correlations with criterion variables deemed socially desirable than the one in which it wasn't included, such as self-perceived attractiveness or intelligence, and even objectively observable behaviours like grade-point average, blood donations, and similar (Bäckström et al., 2014; Bäckström & Björklund, 2013; 2020).

Combating socially desirable responding in personality research

Operating under the premise that social desirability sullies the validity of established personality inventories, researchers began thinking of ways to mitigate its unwanted effects. The most widespread practice is post-data collection statistical control based on social desirability values estimated through specific social desirability scales, the so-called covariate techniques (Paulhus & Vazire, 2007, as cited in Bäckström et al., 2014). One of the first non-pathological, and still widely used, measures of social desirability is the Marlowe-Crowne Social Desirability Scale (MC-SDS) (Crowne & Marlowe, 1960). MC-SDS has, however, garnered many criticisms, with main issues cited being its length and the restriction of response variability as a consequence of its forced-choice format (Lee & Sargeant, 2011). Efforts made towards validation of this scale also raised concerns regarding the unidimensionality of social desirability, prompting a pivot towards scales built upon the assumption of multidimensionality of the construct (Leite & Beretvas, 2005).

Perception of social desirability as composed of two orthogonal factors, *impression management* and *self-deceptive enhancement*, is currently the leading perspective on social

desirability. Of the two, impression management (IM) pertains to the public domain, actively manipulating others' perception of one's image; self-deceptive enhancement (SDE), as the name hints, pertains to the private domain, reflecting one's passive, unaware manipulation of their own perception of self. The Balanced Inventory of Desirable Responding, or BIDR, (Paulhus, 1991) lies upon the basis of these two factors, split evenly into two relatively independent scales (Paulhus, 1991; Pauls & Stemmler, 2002; Paulhus & Trapnell, 2008). Most impression management scales, as does the BIDR one, rest upon the idea that individuals who are inclined to present in a socially desirable way claim highly desirable rare behaviour and deny common, but undesirable behaviour; thereby allowing determination and statistical control of the extent of response bias (Müller & Moshagen, 2018). While Ones et al. (1996) maintain that the pattern of correlations between personality measures and social desirability remains similar regardless of specific subscale, impression management seems to be somewhat more sensitive of the two factors to faking in self-presentation, thus influencing the validity of personality measures (Holden, 2007; Rosse et al., 1998). It was Paulhus (1991) himself who recommended control of IM, but not SDE, in self-report measures, precisely because of its deliberate nature.

Some argue, however, that correcting for social desirability isn't a viable solution to dealing with variance unexplained by the five factors. Concerns were raised after it was repeatedly shown that validity, be it construct, convergent or predictive, of personality scales does not necessarily increase with removal of personality items contaminated by impression management, nor with statistical control of social desirability; on the contrary, some claim that this practice may even result in overcorrection, that is in sacrificing substance of personality rather than removing method variance (Borkenau & Ostendorf, 1992; McCrae & Costa, 1983; Pauls & Stemmler, 2002; Piedmont et al., 2000). In an attempt to avoid adverse consequences of the previously mentioned approach to managing response bias, researchers have turned to prevention rather than an intervention. One of the ways to do this is to either minimise the need for dissimulation by influencing the context of survey completion, such as emphasising the importance of honesty or avoiding using excessively evaluative items altogether (Paunonen & LeBel, 2012). While the former does not guarantee a method variance-free data, the latter seems especially challenging in the case of measures of traits whose valence is obviously observable, as well as raises questions regarding the validity of the scale one is left with after a number of items is forcibly removed. The first to suggest keeping all the personality scale items, but ridding them of their evaluative content altogether,

were Peabody and Goldberg (1989), followed to a degree by Saucier (2002). Contemporary efforts made toward creating a personality inventory free of the evaluative method variance build upon the basis of these researchers and focus on item wording (Bäckström & Björklund, 2013; 2020; Bäckström et al., 2009; 2014; 2023). After extracting the common factor that was causing factor intercorrelations, Bäckström et al. (2009) sought to test whether a neutralised personality inventory based on the existing IPIP-100 formulated by rewording negative items as less negative and positive as less positive would retain the FFM structure while simultaneously being under less influence of social desirability. Analyses showed that IPIP items produced a larger principal component compared to the neutral items (.40 v. .12) and higher mean correlation to social desirability (.37 v. .20) while retaining factor structure, albeit suffering a somewhat lower reliability. The success of this study prompted years of scale development, culminating with the Neutralised Big Five Inventory (NB5I) (Bäckström et al., 2023). The inventory in question evolved from rephrasing of traditional IPIP-100 items by undergraduate psychology students with a method called evaluative neutralisation (Bäckström & Björklund, 2013). The goal was to produce an item pool which altered the popular items, those with a mean above the midpoint of the scale when desirable and those with the mean below the midpoint of the scale when undesirable, keeping their corresponding descriptive content while simultaneously ridding them of evaluative content. Examples of rephrasing using this method produced were, for example, a change from ‘Avoid contacts with others’, which sounds fairly undesirable and even antisocial, to ‘Feel at ease even with being alone’, a statement that is neither desirable nor undesirable; a similar change has been done on items signalling desirability, e.g. ‘Have vivid imagination’ to ‘Constantly spend time exploring my imaginary world’ (Bäckström et al., 2014). The new inventory was created from items with the lowest correlation with social desirability and, similar to their previous studies, that best compared to the original personality inventory. Items loaded onto correct facets and factors, with the exception of Agreeableness facets, and less variance was explained by the principal component — that is, the new inventory contained less common variance and more independent factors. The correlations between the inventories were satisfactory, with the notable exception of Emotional stability. Regardless, subsequent studies confirmed that evaluative neutralisation can be carried out without sacrificing criterion or discriminant validity (Bäckström & Björklund, 2020; Bäckström et al., 2014).

Alternative view of SDR — style versus substance

There is, however, an ongoing debate about whether the common factor, understood as social desirability, represents a stable trait or method variance (Connelly & Chang, 2015; Holden, 2007; Holden & Passey, 2010; McCrae & Costa, 1983). Traditionally, social desirability scores were considered a nuisance factor whose effects were necessarily removed to ensure validity of the scale in question (Connelly & Chang, 2015). As awareness of the risk of removing relevant personality variance or otherwise useful variance rises, especially using covariate techniques of separating the two, this conservative approach has been contested in the last several years (Bäckström et al., 2014). Several points of contention were recognised. For one, rank order of participants appears to resist SDR corrections and does not collapse the factor structure of the Big Five (Linden et al., 2011; Lonnqvist et al., 2007). Moreover, Bäckström and Björklund (2009) show that factor structure remains stable even when the items are weakly related to social desirability, suggesting that social desirability is not a factor that is directly part of the FFM. Secondly, studies regarding the general factor of personality have offered intriguing takes on its role within the social desirability research. It appears that in the comparison of high- and low-stakes situations the factor structure remains stable, and thus do the relative differences in personality variance explained by the general factor of personality modelled at item level (Anglim et al., 2017; Van der Linden et al., 2011). While mean levels in high- compared to low-stakes contexts remained higher on the general factor primarily, but also on basal personality dimensions, its relative consistency across situations should not be dismissed as an argument for the hypothesis of it not being solely a product of response bias. Thirdly, SDR scales have in cases been shown to be indistinguishable from personality measures, though IM and SDE scales showed greater distinction than other measures that were used (Bensch et al., 2017). SD corrections on personality scales have not necessarily improved the scales' personality-related criterion validity either, in some cases even weakening them, removing substance rather than the presumed bias and rejecting the notion of socially desirable responding as a moderator or a suppressor variable in personality criterion prediction (Connelly & Chang, 2015; Ingold et al., 2014; Li & Bagger, 2006; Ones et al., 1996; Pauls & Stemmler, 2002; Piedmont et al., 2000). SDR has been shown to correlate not only with personality traits, presumably tapping into the same processes as neuroticism especially (Bäckström & Björklund, 2013), but also with positive and negative affectivity (Biderman et al., 2011), and predicts antisocial behaviour such as substance abuse (Verschuere et al., 2015). SDE scales related positively to a self-serving bias, self-esteem, illusion of control, and a number of coping/self-defence styles such as repression and distancing (Holden et al., 2000; Paulhus, 1991; Paulhus & Reid,

1991), as well as negatively to depression, emphatic distress, social anxiety and neuroticism in general (Meston et al., 1998). They were also weakly correlated to environmentally relevant variables, ranging from climate change beliefs and intention to sacrifice for the environment to support for collective action to address environmental issues, though only for short-term intention and behaviour, and higher for impression management (Vilar et al., 2020). Self-enhancement bias specifically, observed in relation to a broader social structure, reflected positively on self-esteem and subjective well-being, but negatively on interpersonal adjustment (Kwan et al., 2004). IM scales were confounded with features of personality, and correlated positively with resilience to stress, satisfying interpersonal relationships, morality, dutifulness, and other adaptive traits and life outcomes (Ones et al., 1996; Paulhus & John, 1998; Uziel, 2010, 2013). Impression management also showed greater consistency in self-other report studies, explaining no significant variance between the two types of raters, providing further proof of its trait-like substantive content (Borkenau & Zaltauskas, 2009; Piedmont et al., 2000; Uziel, 2013). Similarly, SD scales have sometimes correlated not only with traits measured by self-reports but also with other-reports, indicating substantive personality variance not entirely bound to response style (Ones et al., 1996; Konstabel et al., 2006). Finally, a subset of researchers argue that the concept of socially desirable responding in itself contains substantive content, taking into account or not further prediction based on the responses. They claim that SDR, namely Impression Management, at least partly represents the subject's knowledge of social norms and their either hypothetical or real alignment with them, or what is otherwise known as cultural normativity and consonance or person-group congruence. In that case, responses on SDR scales would not necessarily reflect the true self and objective behaviour, but perhaps the desired self and values (Bou Malham & Saucier, 2016; Connelly & Chang, 2015).

As result of the above, some personality researchers concluded that SDR is not cause for concern in personality assessment (Borkenau & Zaltauskas, 2009). Most of those agree that is because of SDR's substantive content, with its correlations with personality factors revealing it as part of a broader spectrum of personality, but nevertheless maintain that it should be measured separately (Bäckström et al., 2014). A great number agrees that the construction of an SDR scale that does not capture variance due to individual differences in personality might even be near to impossible (Holden & Passey, 2010). Others, mainly those studying the general factor, are wary of stating whether its socially desirable nature is substantive or artefactual without correlating it to more objective, external criteria (Pelt et al.,

2019). As is often the case, in the end most researchers conclude that the SDR scales measure both, substance like bias, in varying proportion depending on situational context and motivation, and balance unfavourable elements like faking with favourable substantive content (Connelly & Chang, 2015; Bou Malham & Saucier, 2016; Pauls & Stemmler, 2002). With the issue still up in the air, however, it seems the safest approach is that of Nevid (1983), who maintains that social desirability is not to be regarded as a confound unless the covariation in question is theoretically inconsistent or a threat to factorial content. After all, personality scales are susceptible to conscious faking, and we do not treat them as if they measure solely response sets instead of substantive traits; amenability of a scale is an insufficient criterion in rendering its content a response set instead of a standalone trait (de Vries et al., 2014).

Social desirability and honesty

Now that it has been established that the original vision of a construct that has been dubbed social desirability bias may be outdated and its contents, supposing their substantive meaning, beneficial in prediction of behaviour and establishing criterion validity, the nature of said substance ought to be considered.

As was discussed above, theories of researchers accepting the new view of SDR range from accepting it as a “necessary evil”, a distortion of results aligned with personality in a way that improves prediction, to a core personality trait integral to the Big Five framework (Bensch et al., 2017; Connelly & Chang, 2015; Ingold et al., 2014; Pauls & Stemmler, 2002). What is particularly interesting, however, especially considering SDR’s alternative label *faking* (e.g., Christiansen et al., 1994; McFarland et al., 2003) and even the existence of SDR scales perhaps, or perhaps not, fittingly named *Lie Scales* used within Eysenck’s Personality Questionnaire (EPQ) (Eysenck et al., 1985) and The Minnesota Multiphasic Personality Inventory-3 (MMPI-3) (Ben-Porath & Tellegen, 2020), is that few researchers challenge the cemented conviction of SDR’s basal insincerity. It appears the idea of a person responding to a social desirability scale truthfully, as they presumably respond to the personality items themselves, instead of deceitfully, is difficult to imagine. Some researchers even seem especially adamant in using SDR as direct measures of dishonesty, as judged by laypersons (e.g. Feldman, 2018). The lack of resistance to this idea comes as a bit of a surprise, however, since issues with this belief have been raised in the studies of the HEXACO model (Lee & Ashton, 2016), as opposed to the IPIP, either being less susceptible to social desirability bias

or having it more appropriately accounted for (e.g. de Vries, 2010, 2014; 2017; Müller & Moshagen, 2018; Zettler et al., 2014). As the reader is certainly familiar, the main difference between IPIP and HEXACO personality inventories is the addition of the sixth dimension of personality, dubbed Honesty-Humility, which sets dishonesty, fraud, greed, and feelings of self-importance on its low pole in opposition to sincerity, fairness, greed avoidance, and modesty on its positive pole (de Vries et al., 2014). Correlations of both self- and other-rated Honesty-Humility with social desirability measures such as the Impression Management scale, along with its lack of traditionally expected moderation in prediction of real-life outcomes, raised concerns over Impression Management's suitability in measuring response bias (Müller & Moshagen, 2018). Namely, in de Vries et al.'s (2014) study SDE and IM scales of the BIDR correlated with Honesty-Humility dimension of the HEXACO $-.03$ ($p = .76$) and $.44$ ($p < .01$), respectively. Specifically, IM's strongest correlate was the Honesty-Humility facet Fairness; Honesty-Humility facet Sincerity was also significantly related to responses on both self- and other-ratings, $.44$ and $.47$, respectively. Thus, the authors concluded the marrow of a substantive interpretation of IM was this: high scorers on IM are actually more honest/have more integrity than low scorers on IM. A similar conclusion was reached by Zettler et al. (2014), whose study went further and showed high scores on IM not only correlate highly with Honesty-Humility ($r = .55$), but also translate to real-life behaviour; the high-scoring participants engaged in less cheating behaviour in their coin-tossing task than those scoring low. They do, however, stress the importance of low-stake context, such as was their own and most anonymous personality studies, in interpreting the IM scale as an honesty measure. Another argument for re-evaluation of the interpretation of the meaning behind Impression Management scales is a sort of self-other-report correlation reported by Ingold et al. (2014) in which pre-selection self-presentation showed not only a lack of negative correlation, but a positive trend in predicting supervisor-rated performance review, implying possible successful long-term prediction of objective behavioural outcomes based on subjective self-reporting. Additionally, while most studies dealing with the topic at hand focus on the relation of SDR to Honesty-Humility, the rare cases, such as Holtrop et al.'s (2020) think-aloud study, also support the idea by analysing social desirability independently. Recent evidence as such points to the truth behind SDR measures actually being a relaxed personal interpretation of the item content not accounted for by the researchers; this particular study's participants' mental assessment of the options presented and their line of thinking being judged as sincere by a third, independent party.

Although the research around socially desirable responding and its consequences regarding scale validity is far from over, it seems that contemporary researchers are more open to reevaluating traditional beliefs that used to permeate the field. As was relayed above, recent developments in the interpretation of the construct behind the most commonly used response bias scales seem to suggest that there is, at least partially, some truth behind the alleged bias.

Current study

The aim of this study is threefold. First and foremost, the basis of the study needed to be set by confirming the existence of the general evaluative factor in personality research, specifically when using the IPIP-NEO-300 (Maples-Keller et al., 2014), a measure developed on the basis of the NEO PI-R FFM model. In order to do this, the common factor was modelled from IPIP-NEO-300 factor items and the recently developed neutralised NB5I (Bäckström & Björklund, 2020). To cement the factor's meaning, the original personality inventory was correlated to the Paulhus's (1991) BIDR-6 measure of social desirability bias. Following previous findings in evaluative factor research (Bäckström & Björklund, 2013; 2020; Bäckström et al., 2009; 2014; 2023), the extracted evaluative factor is expected to share significant variance with BIDR-6. Thus, the first hypothesis was formulated;

H1: The evaluative factor of personality correlates positively with the measure of social desirability.

Secondly, it was in my interest to confirm not only the evaluative factor's theoretical nature, but also its practical impact on real-life behaviour. Studies showing the effect of social desirability scores on behavioural outcomes such as antisocial behaviour (Verschuere et al., 2015), environmental action (Vilar et al., 2020), and interpersonal adjustment (Kwan et al., 2004) seem to confirm not only the existence of tangible consequences of SDR scores, but also the traditionally unexpected positive relation between any two measures, implying behaviour-accordant self-reporting. These findings led me to construct a new, multiple-response situational scale which was used as the criterion measure to the SDR predictors. Unlike any currently available instrument, the newly-constructed measure, while not a direct measure of behaviour, was intended to be used as an indirect measure of overarching behavioural tendencies based on previous experience. Agreement with aforementioned positions thus informed my second hypothesis;

H2: Both social desirability scores and the evaluative factor will predict adaptive self-reported past behaviour.

Finally, to expand on the substance debate of the SDR research and attempt to answer the question of sincerity of people's responses on SDR measures, a measure of people's self-perceived honesty was observed in relation to the evaluative factor and BIDR data. As was discussed in the previous section, honesty has so far been linked to SDR mainly through HEXACO's Honesty-Humility personality dimension. However, while that scale does contain honesty-related items, not only are there just 4 in the entire 100-item inventory, but the Honesty-Humility dimension also consists of other facets: fairness, greed avoidance and modesty. Moreover, the four items pertaining to sincerity appear to, at least at face value, not only be alarmingly similar to SDR items themselves, thus being either susceptible to it or presumably at risk of having inadequate discriminant validity in that aspect, but also entwined with the morality of item content. To add to the discriminant validity argument, almost half of the BIDR Impression Management items seem to entertain content concerned with integrity more than anything (de Vries et al., 2014). However, while closely related, morality and honesty are not to be mistaken for each other. After all, as not all lies are considered immoral, such as those said to benefit the other, neither are all truths considered moral, such as those said regardless of their impact on others (Levine & Schweitzer, 2014). What is of interest in this study is the justice aspect of morality (Walker & Hennig, 2004) — value of honesty unattached to implied care morality of the acts described. In an effort to separate this construct of trait honesty from HEXACO-relevant morality-based confounds, I ventured to make an instrument that would measure solely people's trait honesty. This scale was then used to test the importance of honesty on SDR scores' prediction of self-reported past behavioural outcomes. The final hypothesis of this study was formed;

H3.0: Honesty moderates the relationship between social desirability and behaviour.

H3.1: High scorers on honesty will exhibit higher positive correlation between social desirability scores and self-reported past behaviour.

H3.2: Low scorers on honesty exhibit lower positive or even negative correlation between social desirability and self-reported past behaviour.

Method

Participants

The participants in this study were recruited through human data-collecting platform Prolific. There were only two requirements for participating in the study: being at least 18 years of age, and living in a country with English as an official language, in order to ensure the high level of fluency in English necessary for understanding the nuances in item wording. After excluding participants who either did not fit the criteria, left the survey entirely empty or, indicating the level of attention below acceptable, responded to basic questions in a nonsensical fashion (e.g. responding to a question requiring the name of a country with the year they were born), a grand total of 35, the final number of participants reached a conclusive $N = 299$ people. Most came from either the US ($N = 192$) or the UK ($N = 86$), with the remaining participants not forming a specific group. The ratio of men to women was 146 to 152, respectively (one person reported their gender as non-binary). The age range of the sample was 18-67 years, with the mean age being 35.7 ($SD = 11.0$). Each participant was monetarily compensated for their efforts.

Materials

The Neutralized Big Five Inventory (NB5I) consists of 120 items distributed on five factors; 24 items per factor, each factor with four facets, six items each. The items of the inventory were developed in many cycles (Bäckström & Björklund, 2020; Bäckström et al., 2014), the final inventory being presented in Bäckström et al. (2023). The ratings on NB5I are made on a 5-point Likert scale ranging from completely disagree to completely agree. The headings of these ratings are formulated as “I.../I am.../I believe that...” followed by the statements. The mean values of many NB5I facets are close to the midpoint of the rating scale (2.00), implying perception of items as neither popular nor unpopular as the neutralisation method intended. Comparing NB5I to the IPIP-NEO, the percentage of mean item ratings above 2.3 was 16.7% for the NB5I as opposed to IPIP-NEO’s 63.5%, suggesting significantly lower evaluativeness of the former. All five factors of NB5I show very high homogeneity, between .85 and .91. As for the general factor, the hierarchical omega was found to be .28 for the NB5I as opposed to .51 for the IPIP-NEO, indicating a substantially stronger general factor in the IPIP-NEO. In fact, the NB5I model does not suggest a general factor at all. Therefore, the NB5I is generally evaluatively neutral, its factors reliable, normally distributed, and independent (aside from Extraversion and Openness), and the facets without strong secondary loadings. The NB5I has been shown to be more appropriate for a simple structure model as compared to IPIP-NEO.

IPIP-50. The International Personality Item Pool-NEO Personality Inventory (IPIP-NEO-300) is based on the NEO-PI-R (Costa & McCrae, 1992) and has 300 items distributed on five factors: Extraversion, Agreeableness, Conscientiousness, Emotional stability and Openness; with six facets each. The ratings are made on a 5-point Likert scale ranging from completely disagree to completely agree. As it has been widely used for quite some time already, IPIP inventories' validities and reliabilities have been duly tried and confirmed, among others by Maples et al. (2014). According to their testing, IPIP-NEO-300's domain score alphas range from .91 to .94, with a median of .93; while alphas for the facet scores range from .58 to .88, with a median of .82. At domain level for the 300-item IPIP-NEO, convergent correlations range from .88 to .91, with a median of .89, while facet-level convergent correlations range from .55 to .85, with a median of .76. The overall mean convergent correlation for the IPIP-NEO scores is .78. Discriminant validity correlations range from .39 to .21, with an absolute median correlation of .17. As for internal consistency, scores from the NEO PI-R demonstrated mean coefficient alphas of .91 for the domains and .82 for the facets, respectively. It is conclusive that this pool is well validated and, with good reason, widely used in the assessment of the FFM. The 50 items used in this study retained factor structure and showed scale level Cronbach's alpha internal consistency of .774.

The Balanced Inventory of Desirable Responding (BIDR-6) —Version 6; Form 40 (Paulhus, 1991), is a descendant of the Self- and Other-Deception Questionnaires developed by Sackeim & Gur (1978) and modified from the Balanced Inventory of Desirable Responding (Paulhus, 1988). The 40-item BIDR measures socially desirable responding across two constructs: Self-deceptive Enhancement (20 items; e.g., "My first impressions of people usually turn out to be right") and Impression Management (20 items; e.g., "I sometimes tell lies if I have to"). Respondents rate their agreement with each statement on a 7-point Likert scale. As one of the most used measures of socially desirable responding, the validity of the BIDR 6 is supported by extensive research. Early research by Paulhus himself established strong discriminant, convergent, and construct validity. The reliability of the BIDR 6 subscales is generally robust; Cronbach's alpha typically ranges between .75 and .86 for both subscales across diverse populations, indicating acceptable internal consistency and homogeneity.

Situations Scale. To measure self-reported past behaviour, a scale measuring retrospective behaviour was constructed specifically for the needs of this study. The scale

consists of 20 relatively commonly experienced situations, each one followed by five possible responses. The responses are formulated so that they vary in how actively or passively socially desirable or undesirable they are. (e.g. *You made a mistake. a) I ignored it* (passively undesirable), *b) I accepted the consequences* (neutral), *c) I fixed the mistake as quickly as possible* (passively desirable), *d) I put the blame on someone else* (actively undesirable), *e) I took responsibility and made sure it wouldn't happen again.* (actively desirable)). The participants are instructed to recall a time in the past month when they experienced a situation as described and choose the one option that best reflected how they responded, and one that reflected their actual response the least. The intent of the scale is to encompass as many as possible different situations, while still leaving them relatively vague to ensure the situations are relatively general experiences. The scale was formed through discussions with several colleagues and friends, and modifications were made according to feedback to ensure comprehensibility and reduce overlap between the possible answers. For the full scale, look at Appendix A.

Honesty Scale. Due to the lack of a validated instrument measuring exclusively trait honesty, I constructed a special 21-item scale exploring this construct. Considering the survey nature of the present study, and thus in order to maximise predictive validity of actual behaviour in accordance with this indirect measure of honesty, it was developed with Ajzen's (1991) Theory of Planned Behaviour in mind. The scale was thus separated in three parts, each containing 7 items: Attitudes (e.g. *I believe honesty is one of the most important traits a person can have.*), Norms (e.g. *Because people around me are honest, they expect me to be honest too.*), and Perceived Behavioural Control (e.g. *Telling the truth or lying is completely under my control.*). The combination of these three aspects is meant to ensure successful prediction of actual honest behaviour based on the provided responses. The participants are instructed to rate their answers on a 5-point Likert scale, ranging from 1 - Strongly Disagree to 5 - Strongly Agree. Provided reversal of the negative items was done, higher score on the scale would signify greater evaluation and practicing of honesty, while lower scores would signify lesser evaluation and practicing of the trait of honesty. For the full scale, look at Appendix B.

Procedure

The study was conducted online, in June of 2025. The participants accessed the survey as registered users in Prolific. Once they clicked on the survey, they were given a

vague description of the study without mention of social desirability, but emphasising anonymity, voluntariness and *honesty*. They were reminded that they were free to exit the survey at any time. The participants then had to confirm that they have read the instructions and wished to proceed with the survey. The instrument battery was presented in the following order: NB5I and IPIP-NEO items in random order, BIDR-6, Honesty Scale, and finally Situations Scale. All the items within these broad separations in the instrument battery were presented in randomised order. To reduce fatigue, help retain attentional focus and increase motivation by shortening the survey, only 50 of the original 300 items from the IPIP were incorporated. Additionally, each participant was randomly assigned only 7 out of 20 possible situations from the Situation test, as well as presented with only half the available items from both the NB5I and BIDR.

Ethical considerations

This study is in compliance with the ethical guidelines outlined by the Law (2003:460) on Ethics of Research Involving Humans (<https://etikprovningssmyndigheten.se/>). All participants were given a description of the research they were contributing to, prior to participating. They were assured that they may, at any point during the survey, withdraw their consent by exiting the survey before submission. After the information was read, participants had to explicitly consent to participation by continuing to the actual survey. Neither sensitive data nor data which can be used to trace or identify specific people other than age, nationality and gender was collected. The participants were reassured that the collected data was not to be used for any other purposes other than academic research. The survey did not pose a risk to either mental or physical well-being of the participants. No participant was underage.

Results

Analysis

The hypotheses were tested using the statistical programme Jamovi (The jamovi project, 2026) using the Regression, Factor and medmod packages. Data was prepared using Microsoft Excel (2025), reverse-coded where necessary, and cleared of aberrant results.

The design for this study was five separate instruments; one of which was a 21-item scale (Honesty Scale) all to which all participants responded to, one of which (IPIP-NEO) contained around a fifth of its full item pool but to whose presented items participants responded to in entirety, and three of which (NB5I, BIDR-6, and Situations Scale)

participants were randomly assigned to respond to 50%, 50% and 35% of the full presented item set, respectively. The remaining missing data was estimated accordingly; the psychometric approach planned missing data was taken to ensure the validity of each participant's data even on reduced measurements. In multivariate planned missingness such as this one, where each participant completes only parts instead of the full extent of the questionnaires, an estimation of the full response set, different for different participants, is achieved using full information maximum likelihood (FIML) under assumption of data *missing completely at random* (MCAR) (Graham et al. 2006).

Evaluative factor and social desirability

The participants were instructed to fill out both an evaluative personality inventory (IPIP-NEO) with items containing evaluative subtext along with standard content and a neutralised version based on the evaluative inventory with the evaluative subtext eliminated (NB5I). For a comparison of descriptive statistics of the two inventories, refer to Table 1.

A paired-samples t-test was used to compare the differences between mean response values to the Big Five domains on the two inventories. All five personality factors showed highly significant ($p < .001$) mean differences in responses to the two inventories. The largest differences were noted for Conscientiousness ($d = 1.370$) and Agreeableness ($d = 1.285$), while the lowest was reported for Emotional stability ($d = .312$; the only small effect). Observing mean responses to the Likert scales, it is clear that people tend to rate the neutralised items lower, closer to the midpoint of the Likert rating scale, than they do evaluative items. For more, look to Table 2.

Table 1

Comparison of descriptive statistics of the evaluative and neutralised personality inventories.

	IPIP			NB5I		
	<i>M</i>	<i>SD</i>	<i>SE</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
Extraversion	3.49	.838	.049	3.10	.598	.035
Agreeableness	3.98	.631	.037	3.16	.527	.031
Conscientiousness	4.05	.537	.031	3.36	.558	.032
Openness	3.90	.577	.033	3.43	.588	.034

Emotional stability	3.17	.706	.041	3.02	.635	.036
---------------------	------	------	------	------	------	------

Table 2

Means, standard deviations, pair-wise t-test and p values and Cohen's d effect size for the difference between IPIP and neutral personality domains.

	Mean diff.	Standard Deviation diff.	<i>t</i>	<i>p</i>	<i>d</i>
Extraversion	.389	.0380	10.24	< .001	.592
Agreeableness	.819	.0369	22.21	< .001	1.285
Conscientiousness	.689	.0291	23.69	< .001	1.370
Openness	.479	.0296	15.88	< .001	.919
Emotional stability	.152	.0281	5.40	< .001	.312

Table 3 shows Pearson correlations between the five personality domains measured by IPIP-NEO and the NB5I. As is evident, most correlations were significant, the ones between the relevant IPIP-NB5I pairs all in the moderate bracket, ranging from $r = .405$ for Agreeableness, which was the lowest recorded coefficient, to $r = .742$ for Emotional Stability, which was the highest recorded coefficient. It is worth noting that while intercorrelations of the factors within the neutralised inventory are not all significant and are overall much lower than the intercorrelations within the IPIP, they have, at least not in this sample, been completely eliminated. Overall, however, the NB5I appears to be less evaluative and closer to orthogonality of the five factors, while maintaining factor structure, compared to the original inventory.

Table 3

Pearson Correlations for the FFM dimensions measured by IPIP-NEO and NB5I.

	E _n	A _n	C _n	O _n	ES _n	E _e	A _e	C _e	O _e	ES _e
E _n										
A _n	.184**									
C _n	.337***	.190***								
O _n	.285***	.134*	.437***							
ES _n	.039	-.006	.166**	.050						
E _e	.626***	.114*	.368***	.380***	.379***					
A _e	.268***	.405***	.294***	.422***	.139*	.452***				
C _e	.253***	.178**	.579***	.420***	.288***	.507***	.538***			
O _e	.179**	.132*	.350***	.614***	.179**	.472***	.547***	.578***		
ES _e	.061	.048	.207***	.080	.742***	.362***	.233***	.358***	.228***	

Note. * = $p < .05$, ** = $p < .01$, *** = $p < .001$. E = Extraversion, A = Agreeableness, C = Conscientiousness, O = Openness, ES = Emotional Stability. _n = NB5I, _e = IPIP

Finally, now that the basis for it has been set, to test the first hypothesis the principal component, in this case referred to as the evaluative factor, was calculated by subtracting the mean values of the NB5I inventory dimensions from the mean values of the original IPIP values for each participant separately. This newly-formed variable was then standardised and subsequently correlated on scale-level with the standardised measures of social desirability, the subscales of the Balanced Inventory of Desirable Responding separated. Both Self-deceptive Enhancement and Impression Management correlated positively and significantly with the data-extracted evaluative factor. It is necessary to point out that both relationships were relatively weak, especially compared to previous findings, though it is worth noting that the correlation with the evaluative factor was slightly larger in the case of SDE compared to IM; $r = .368$ and $r = .186$, respectively. Thus, the evaluative factor does correlate positively with measures of social desirability, although the BIDR subscales together explain only less than 15% of the entire evaluative factor variance. The correlation matrix can be found in Table 4, below.

Table 4

Correlation between evaluative factor and Balanced Inventory of Desirable Responding subscales.

	EF	BIDR (SDE)	BIDR (IM)
EF			
BIDR (SDE)	.368***		
BIDR (IM)	.186**	.389***	

Note. * = $p < .05$, ** = $p < .01$, *** = $p < .001$. EF = evaluative factor, BIDR (SDE) = Balanced Inventory of Desirable Responding (Self-deceptive Enhancement), BIDR (IM) = Balanced Inventory of Desirable Responding (Impression Management)

Predicting behaviour with Situations Scale

To examine the second hypothesis, I turned to the newly-constructed measure of self-reported past behaviour. To encode the responses optimally, I've chosen to calculate each participant's Situations Scale score by multiplying their selections for the situation they chose, the one fitting their actual response the least as -1 and the one fitting their actual response the most as +1, with each desirability value of the chosen response to a situation, ranging from 1 to 5. This variable was then standardised and ultimately used as the criterion in a regression analysis with the subscales of the BIDR entered into the equation as the two predictors. A confirmatory factor analysis was performed prior to further analyses. Item number 20 was excluded from the analyses due to missing values across all participants. Factor loadings were overall moderately large, ranging from .399 to .726. The complete results of the CFA are visible in Table 5. The model demonstrated a chi-squared test value of, $\chi^2(152) = 304$, $p < .001$, its significance indicating poor fit. As χ^2 index can often be unreliable, other indices were consulted. The model showed mixed evidence, poor fit as reported by all indices (CFI = .721, TLI = .686, 90% CI [.049, .068], and SRMR = .145) aside from RMSEA = .058. In addition, reliability analyses were employed. The scale demonstrated high internal consistency, with Cronbach's $\alpha = .906$ and McDonald's ω estimates ranging from .882 to .907.

Table 5

Factor loadings for items of the Situations Scale.

Indicator	Estimate	SE	95% Confidence Interval		Z	p	β
			Lower	Upper			
1	.190	.036	.119	.261	5.22	<.001	.552
2	.290	.061	.169	.411	4.69	<.001	.582
3	.226	.052	.123	.328	4.31	<.001	.461
4	.207	.052	.105	.309	3.99	<.001	.399
5	.191	.041	.111	.271	4.68	<.001	.473
6	.270	.038	.196	.344	7.12	<.001	.662
7	.279	.050	.180	.378	5.53	<.001	.561
8	.282	.052	.181	.384	5.44	<.001	.547
9	.329	.041	.248	.411	7.95	<.001	.701
10	.184	.033	.119	.249	5.54	<.001	.594
11	.279	.037	.206	.351	7.55	<.001	.665
12	.276	.057	.165	.388	4.87	<.001	.518
13	.254	.035	.187	.322	7.37	<.001	.689
14	.364	.044	.279	.450	8.34	<.001	.714
15	.267	.050	.168	.365	5.29	<.001	.507
16	.336	.042	.253	.418	7.98	<.001	.726
17	.281	.043	.198	.365	6.60	<.001	.671
18	.196	.038	.121	.271	5.13	<.001	.558
19	.243	.046	.154	.333	5.32	<.001	.596

The multiple linear regression was conducted to examine whether Impression Management (IM) and Self-Deceptive Enhancement (SDE) predicted self-reported socially desirable past behaviour. The main fit index assumed 11.5% of variance in self-reported socially desirable behaviour explained by socially desirable responding to the personality measure ($R^2 = .115$), and a closer look showed that the overall regression model was significant, $F(2, 295) = 19.2$, $p < .001$, with even the adjusted $R^2 = .109$. It is visible in Table 6 that both of the two BIDR subscales were significant predictors of actual behaviour on their own.

Table 6

Multiple linear regression model with BIDR subscales as the predictors and behaviour operationalised through the Situations Scale as the criterion.

Predictor	Estimate	β	SE	T	p
Intercept	.560		.017	33.22	<.001
BIDR (SDE)	.069	.222	.018	3.74	<.001
BIDR (IM)	.057	.184	.018	3.10	.002

Linear regression was also employed to test the potential of the evaluative factor to predict socially desirable behaviour and was similarly successful with the model fit index R^2 reaching .051 and the adjusted $R^2 = .048$; overall regression model was significant, $F(1, 297) = 16.1, p < .001$. Full regression model can be seen in Table 7. In conclusion, both social desirability scores and the evaluative factor, to varying degrees, successfully predicted adaptive past behaviour.

Table 7

Linear regression predicting behaviour based on the evaluative factor.

Predictor	Estimate	β	SE	T	P
Intercept	.560		.017	32.21	<.001
EF	.070	.227	.017	4.01	<.001

Honesty and moderation

The newly constructed Honesty Scale was subjected to a confirmatory factor analysis to analyse its intended three-part structure. The model specified three correlated latent factors, with seven items loading onto each factor. All of the item factor loadings were significant and their standard estimates ranged from mid to high, with the exception of two items whose loadings failed to reach values above .200 and had to be excluded from the

analysis. The items in question, number 17 (*No one can stop me from lying if I want to*) and 20 (*Sometimes, circumstances make it hard for me to be honest*) both belonged to the Social Control subscale. Item number 21, '*It's not just up to me whether to be honest or lie*', also belonging to Social Control, was kept in the analysis even though its loading (.259) was below what would usually be appropriate, considering the scale is yet new. After the decision to exclude the two inappropriate items, the model fit was tested. The present model demonstrated a chi-squared test value of, $\chi^2(149) = 420, p < .001$, its significance indicating poor fit. However, as χ^2 index can often be unreliable, other indices were consulted. The model demonstrated mixed fit, from poor to somewhat acceptable values of fit to data; CFI = .855, TLI = .834, RMSEA = .078, 90% CI [.069, .087], and SRMR = .063. Thus, whether the CFA can be considered successful, having partially confirmed model fit for this new scale, is uncertain, especially taking into consideration the fact that they are measured without "error" and that there is a considerable amount of unique information in the scale itself. All factor loadings can be found in Table 8. However, the three latent factors were significantly and positively correlated. The strongest association was observed between Norms and Control ($r = .792$), followed by the correlation between Attitudes and Norms ($r = .738$), and finally Attitudes and Control ($r = .661$). High factor intercorrelations, shown in Table 9, while conceptually reasonable, bring the distinction between the three factors into question.

Taking into account the significant overlap between the factors, reliability was calculated as internal consistency both on item and scale level. Scale level Cronbach's alpha showed acceptable reliability of .861. As for item level, both indices, Cronbach's alpha and McDonald's omega, exceeded .80 — largely considered good reliability values exceeding the widely accepted threshold. Results of item reliability analysis are available in Table 10.

Table 8*Factor loadings for items of the Honesty Scale.*

Factor	Indicator	Estimate	SE	95% Confidence Interval		Z	p	B
				Lower	Upper			
Attitudes	Hon_Q7_1	.528	.037	.455	.601	14.15	< .001	.745
	Hon_Q7_2	.595	.039	.518	.672	15.09	< .001	.780
	Hon_Q7_3	.601	.048	.507	.695	12.54	< .001	.682
	Hon_Q7_4	.494	.059	.379	.609	8.44	< .001	.492
	Hon_Q7_5R	.528	.071	.389	.668	7.42	< .001	.442
	Hon_Q7_6R	.527	.074	.382	.672	7.14	< .001	.429
	Hon_Q7_7	.688	.053	.585	.792	13.01	< .001	.700
Norms	Hon_Q7_8	.611	.051	.510	.712	11.90	< .001	.651
	Hon_Q7_9	.475	.061	.355	.596	7.76	< .001	.456
	Hon_Q7_10	.630	.051	.530	.731	12.29	< .001	.667
	Hon_Q7_11	.528	.040	.450	.605	13.33	< .001	.710
	Hon_Q7_12	.495	.0415	.414	.577	11.92	< .001	.652
	Hon_Q7_13	.531	.0548	.423	.638	9.68	< .001	.557
	Hon_Q7_14	.547	.0377	.473	.621	14.52	< .001	.755
Control	Hon_Q7_15	.482	.0629	.359	.605	7.66	< .001	.490
	Hon_Q7_16	.558	.0460	.468	.649	12.14	< .001	.709
	Hon_Q7_17	.190	.0831	.027	.353	2.28	.022	.156
	Hon_Q7_18	.550	.0585	.435	.665	9.40	< .001	.566
	Hon_Q7_19	.414	.0509	.314	.514	8.14	< .001	.519
	Hon_Q7_20R	.183	.0757	.035	.331	2.42	.016	.161
	Hon_Q7_21R	.336	.0849	.170	.503	3.96	< .001	.259

Table 9*Factor covariances for the Honesty Scale.*

		95% Confidence Interval						
		Estimate	<i>SE</i>	Lower	Upper	<i>Z</i>	<i>p</i>	β
Attitudes	Attitudes	1.000 ^a						
	Norms	.738	.039	.662	.814	19.1	< .001	.738
	Control	.661	.055	.553	.768	12.1	< .001	.661
Norms	Norms	1.000 ^a						
	Control	.792	.043	.707	.877	18.3	< .001	.792
Control	Control	1.000 ^a						

Note. ^a = fixed parameter

Table 10*Item reliability statistics for Honesty Scale.*

	If item dropped	
	Cronbach's α	McDonald's ω
Hon_Q7_1	.850	.871
Hon_Q7_2	.850	.872
Hon_Q7_3	.852	.875
Hon_Q7_4	.855	.878
Hon_Q7_5R	.859	.880
Hon_Q7_6R	.860	.881
Hon_Q7_7	.849	.872
Hon_Q7_8	.851	.874
Hon_Q7_9	.857	.879
Hon_Q7_10	.850	.873
Hon_Q7_11	.850	.871
Hon_Q7_12	.851	.873
Hon_Q7_13	.855	.877
Hon_Q7_14	.849	.870
Hon_Q7_15	.858	.880
Hon_Q7_16	.850	.873
Hon_Q7_18	.855	.878
Hon_Q7_19	.857	.880
Hon_Q7_21R	.871	.886

The honesty scale was subsequently correlated to measures of Impression Management, Self-deceptive Enhancement, and the evaluative factor to paint a better picture of the relationships within this framework. The correlation matrix can be seen in Table 11. The values clearly show and expected, moderate, positive and highly significant, association between all relevant measures entered into the regression, indicating that participants scoring

highly on measures of honesty also tend to score higher on social desirability measures and the evaluative factor.

Table 11

Correlations of Honesty with BIDR subscales and the evaluative factor.

	Honesty	BIDR (SDE)	BIDR (IM)	EF
Honesty				
BIDR (SDE)	.319***			
BIDR (IM)	.386***	.389***		
EF	.322***	.368***	.186**	

Note. * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

Proceeding with the Honesty Scale now that its factor structure and internal consistency were thoroughly examined, the standardised variable calculated based on participant responses to the scale was entered as a moderator in a moderation analysis in order to attend to the third hypothesis. The moderation analysis was conducted to examine whether honesty moderated the relationship between social desirability and self-reported socially desirable past behaviour. Honesty, while affirming significant main effects of honesty and of the two out of three observed constructs on self-reported behaviour, unfortunately failed to show significant moderation of honesty on the relationships in all of the three presented cases; most notably it failed to moderate the relationship between Impression Management, that was most expected to be related to honesty based on past research as relayed prior in the thesis, and self-reported past behaviour. Detailed information on these analyses can be found in Tables 12, 13 and 14.

Table 12*Moderation estimates for BIDR-SDR.*

	Estimate	SE	Z	P
BIDR (SDE)	.058	.016	3.566	<.001
Honesty	.108	.016	6.798	<.001
BIDR (SDE) * Honesty	-.009	.015	-.609	.542

The interaction between BIDR-SDR and honesty was not significant, $b = -.009$, $SE = .015$, $z = -.609$, $p = .542$, indicating that honesty did not significantly moderate the association between self-deceptive enhancement and self-reported socially desirable past behaviour.

Table 13*Moderation estimates for BIDR-IM.*

	Estimate	SE	z	P
BIDR (IM)	.041	.016	2.561	.01
Honesty	.112	.016	6.948	<.001
BIDR (IM) * Honesty	-.012	.015	-.777	.437

The interaction between BIDR-IM and honesty was not significant, $b = -.012$, $SE = .015$, $z = -.777$, $p = .437$, indicating that honesty did not significantly moderate the association between impression management and self-reported socially desirable past behaviour.

Table 14*Moderation estimates for the evaluative factor.*

	Estimate	SE	Z	P
EF	.031	.016	1.885	.059
Honesty	.116	.016	7.221	<.001
EF * Honesty	-.008	.015	-.526	.559

The interaction between the evaluative factor and honesty was not significant, $b = .008$, $SE = .015$, $z = -.526$, $p = .559$, indicating that honesty did not significantly moderate the association between the evaluative factor and self-reported socially desirable past behaviour.

The third hypothesis was not supported; while associated with both measures of social desirability and the evaluative factor partially explained by SDR, and majorly significantly predicted variance in self-reported behaviour, honesty did not in any way moderate the relationship between scores on social desirability scales and self-reported accordant past behaviour.

Discussion

The hypotheses of the current study were only partially supported. As was expected, I managed to replicate the findings regarding the evaluative factor in studies comparing the evaluative IPIP items to neutralised items based on the same (Bäckström & Björklund, 2013; 2020; Bäckström et al., 2009; 2014; 2023). The neutralised NB5I personality inventory exhibited lower average ratings compared to the original non-neutralised items on all five dimensions, implying lower susceptibility to evaluativeness and popularity of items. In this aspect, the most successful neutralisation was done on Conscientiousness and Agreeableness, while the least successful was Emotional stability, and is as such aligned with previous research. This particular finding is hardly surprising, taking into account the content of this specific factor – with facets like Anxiety, Anger, and Depression, it is hardly an easy feat to

separate evaluativeness from core item substance, rendering it comparatively resistant to evaluative neutralisation. Regardless of the low size effect, the significant difference between the five factors of the two relevant scales remains – and with it the success of the neutralisation method, and subsequently of the NB5I personality inventory. It is also worth noting that the correlations between the relevant pairs of factors of the two inventories, the evaluative and the neutralised, were satisfactory – again, the highest correlation is reported for the Emotional Stability/Neuroticism pairing, indicating perhaps the inability to sever this personality trait from the threat of its social implications. Intercorrelations, on the other hand, show that, at least in this sample, the NB5I is still a far cry from the dream of complete orthogonality of the FFM. Even so, generally lower intercorrelations and their less frequent significant p-values compared to their evaluative counterparts show promise for this new inventory and its prospects.

The comparison of the two inventories made it possible to extract the principal component, also known as the general factor or, in this case, the evaluative factor, presumably causing the intercorrelations of the factors within the original inventory. However, while simultaneously confirming the first hypothesis, the relatively weak correlation between the evaluative factor and the measures of social desirability leaves little to interpretation and much to imagination. One possible explanation lies in the perhaps overly simple subtraction method used to extract the general factor. If more advanced models were used to estimate it, as were used in plenty of prior research, the relationship would perhaps be more pronounced. Still, with less than a sixth of the variance explained by BIDR, the question of the true substance of the evaluative factor remains, at least in this paper, unanswered. What of the remaining variance? Further research is necessary to determine the full content of the item-extracted factor. Additionally, contrary to expectations, Self-enhancement alone accounted for the majority of the explained variance, even though previously mentioned studies hinted at Impression Management being more susceptible to faking. It seems that, after all, anonymous low-stake contexts such as the present one fail to fully activate that self-presentational mechanism.

The second hypothesis was also supported. Both the scores on the BIDR subscales, as well as the evaluative factor, were able to predict self-reported socially desirable past behaviour. It seems that socially desirable *behaviour* is significantly dependent on socially desirable *responding*. This finding opposes the one from Bäckström and Björklund's (2021) experience sampling study which, using an analogous method to predict consistent behaviour

congruent with the Big Five with the evaluative factor, concluded that there was largely no correlation between the two. Supposing that the items on the measure of past behaviour were assessed by participants in good faith, this discovery would challenge the well-ingrained idea that SDR measures, in this case namely the Balanced Inventory of Desirable Responding, account for response bias, as well as dispute its synonymisation with faking or lying, and its subsequent use as a measure of dishonesty. Taking into account its role in previously reported criterion validity of personality scores (Kwan et al., 2004; Verschuere et al., 2015; Vilar et al., 2020) and the HEXACO studies (de Vries, 2010, 2014; 2017; Lee & Ashton, 2016; Müller & Moshagen, 2018; Zettler et al., 2014), the support shown for the second hypothesis in this study continues to build upon the line of research standing for predictive importance of so-called social desirability bias. Additionally, the evaluative factor, whose variance was only partially explained by the construct behind BIDR, also significantly predicted self-reported past behaviour. This finding further supports the importance of social desirability bias as explained previously, but also enforces the idea of specific personality, rather than method, variance causing intercorrelations within the factors — urging future researchers not to automatically label it as method bias and discard it as a result of this preemptive conclusion.

Scores on the Honesty Scale have not moderated the relationship between the three constructs used to predict participant behaviour; although honesty explained a significant part of variance of both BIDR and the principal component evaluative factor, whether someone was more honest or less honest had no effect on the strength of prediction of behaviour based on pertinent self-reporting. Therefore, the third hypothesis was rejected. It is important to note, however, that main effects of honesty throughout all three observed relationships, as well as the main effects of Impression Management and Self-Deceptive Enhancement, were significant. Both honesty and social desirability measures predicted behaviour on their own, regardless of each other. Their interaction, however insignificant, trended negatively. This particular trend implies lower association between self-reported responses and behaviour for people scoring higher on honesty. People presenting themselves more favourably tended to, according to trend, report less socially desirable past behaviour if they placed greater value on honesty, as expected. Therefore, this could be taken to mean people who valued honesty higher tended to be more realistic in their evaluation of their real-behaviour, despite overall considering themselves socially desirable. Moreover, the Honesty Scale itself seems to be relatively successfully constructed, maintaining both the majority of appropriate factor loadings and a somewhat stable factor structure, as well as high reliability. This is particularly

valuable, considering the current lack of an extensive, validated measure of the pure trait honesty construct in the field, with most of the available ones containing more latent constructs than just honesty itself. Two items, item no.17 (*No one can stop me from lying if I want to*) and item no.20 (*Sometimes, circumstances make it hard for me to be honest*) both belonged to the Social Control subscale but had to be excluded due to insufficient factor loadings. To maintain balance within the entire Scale, the scale needs to be revised, and two new items are to be added to the Social Control subscale. However, even with the current rudimentary form, the scale shows promise, especially keeping in mind that it was constructed under Ajzen's (1991) Theory of Planned Behaviour. It is especially not to be discarded as a valid predictor of behaviour. For one, it presently predicted behaviour not measured directly but retrospectively and via self-report, even with the behaviour that *was* measured not entirely closely related to honesty in its pure form; rather, it was used as an adjacent construct to the actual criterion. It is likely that the Honesty Scale would have even more successfully predicted actually honest behaviour, especially had it been tested in more experimental conditions rather than just in a correlational study.

Limitations

There are several limitations this study has encountered. The most prominent one, out of the five measures used, two were constructed just for the purposes of this study. There have been no other studies using these exact scales and thus there is no other psychometric data on reliability or construct, discriminant, divergent, convergent or, most importantly for this study, criterion validity on other data samples. Even here, I haven't fully validated the new scales used as it would exceed the scope of the study at hand, which makes the data analysed, and most of the conclusions drawn from it, largely an educated guess at the underlying constructs. Such scales are also inherently imperfect. The Situations Scale, for example, is cause for greatest concern in this aspect. Disregarding the obvious lack of empirical validation, which would be especially useful had the scale been proven to actually relate to or even predict actual behaviour, some aspects can be criticised even at face value. For one, the scale is relying on a month-long retrospective. There are two possible issues with this; scales requiring remembering, especially emotionally charged situations, are bound to be at risk of memory distortion. Additionally, while incorporated to mitigate the former negative by-product, putting a month-long limit may have restricted participants' relatability to the items, even though the scale items were intended to represent fairly common situations anyone might encounter. '*You made a mistake*' is a fairly likely situation for a person to find

themselves in in a month's time, but '*A friend let it slip they had a private problem*' is more likely an event to happen on a less frequent basis. When one is presented with a situation they haven't come across, but are asked to make a judgment on how they reacted, it is likely they simply made their hypothetical reaction up instead of recalling a similar situation they have found themselves in. It is also quite time-consuming and arduous, the principle for filling out such a scale as confusing for the participant as it is difficult to grade and code for the researcher. The Honesty Scale, while showing promise, is still green — two items had to be excluded, the third was retained purely because of the newness of the scale, and some of the other loadings are still too low for a scale that is being used as if it is validated. As such these scales, at least for now, also depend on this particular sample. Furthermore, issues regarding evaluativeness of the Honesty Scale itself can be also be raised. However, despite the desirability of the responses being relatively evident at face-value, I would argue that a subjective self-report measure of a straightforward construct such as honesty would be hard if not impossible to be attained with the neutralisation method as in the case of IPIP and NB5I. I thus conclude the Honesty Scale has been appropriately constructed and stand behind the call for further development on its basis as it currently is.

The sample of this study is another one of the limitations. The first issue with it has to do with the nature of participants. Despite being monetarily compensated, participants who are regular participants in psychological research such as are registered users on Prolific tend to sometimes respond carelessly in order to finish the survey more quickly, impacting the quality of data provided (Bäckström et al., 2023). The second issue stems from the fact that, while a nearly 300 participant sample is an ample number, and the gender ratio satisfactory, there remains a question of how the results are affected by a such age-diverse sample, considering the difference between the oldest and the youngest participant was almost half a century. This is especially important with a variable such as social desirability, which depends on social norms, which in turn may vary between cohorts. Although explicit care was taken in defining the sample beforehand to avoid bringing cultural variance into the equation, it is also possible, however unlikely, that some of it was introduced despite our best efforts to reduce it. However, as the sample produced satisfying results in testing the first hypothesis, although generating lower correlations than expected based on previous findings, I have to conclude that any issues stemming from the sample in this particular case are negligible.

Another, perhaps obvious, limitation is that the study's aim was to examine the predictive ability of social desirability scales on socially desirable behaviour. However, no

direct measure of behaviour was used in the study, as it was conducted online with just the instrument battery. The relative inappropriateness of such a study design in answering this particular research question severely affects the extent of conclusions that are possible to be drawn from data at present. Caution needs to be exercised in interpreting the responses to retrospective self-report measures of behaviour as they cannot be regarded as interchangeable with actual behavioural measures; their content too unidimensional and relying greatly on both fickle memory and participant integrity in responding truthfully. Any interpretations made from them are therefore restricted by that fact.

Future directions

Partial support for the hypotheses in this study should not deter from future research in the field, especially taking into account all the previous studies that paved the way for redefinition of the concept of socially desirable responding.

There are two major possible future directions I wish to point to. One of them is certainly further development of the Honesty Scale, as it seems to have been the most successful part of the current study. Already promising, if developed fully, the scale may open up the possibility of examining how much an individual truly values honesty in their life. It could impact how we interact with people on a personal basis, but also how we interpret their test scores in high-stake contexts. Theoretical, more advanced, future studies using the scale in question may indeed point towards certainty of sincerity of social desirability scores as well as other responses in general, using it not only in social desirability research, but also other branches of psychology such as clinical psychology, work psychology, or even forensic psychology. Used in these areas, it could help discern between people who are faking and people who are being sincere in contexts with tangible repercussions such as (in)accurate medical diagnoses, (missed) offered job positions, or (in)justly ordered sentences.

The second direction I recommend. for future personality researchers interested in the style vs. substance debate of social desirability specifically, is perhaps disregarding trait honesty completely and instead focusing on trait morality. The HEXACO Honesty-Humility correlations with socially desirable responding point, due to their very nature, in multiple directions. Taking the results of this study into account, one might turn towards morality next, since it seems to be underlying both the entire domain throughout all four facets, including sincerity, as well as the BIDR itself. As was previously emphasised, almost half of the BIDR Impression Management subscale specifically seems to entertain content concerned with

integrity (de Vries et al., 2014). In that case, it would also be of interest to test the differences between trait and state morality, depending on how high- or low-stakes the context in which socially desirable responding is tested, as it seems relatively relevant for the size of social desirability effect itself, as well. This experimental design is one more way of improving the current study, to better clear the view of the differences between self-enhancement and impression management that supposedly depend on, among else, motivation and witnesses.

Conclusion

The present thesis examined and confirmed the existence of the evaluative factor causing factor intercorrelations within the traditional IPIP personality inventory, associated moderately with social desirability as measured by the two-factored BIDR. It also confirmed the ability to predict socially desirable behaviour based on socially desirable responding. However, it failed to find the moderation effect of honesty on that same relationship. While unsuccessful in identifying the substance of the evaluative factor not explained by social desirability and complete social desirability itself, further research is left with a solid basis in the field of honesty, with the newly-constructed and partially validated Honesty Scale, while social desirability research is steered in an adjacent direction based on other recently reported HEXACO-SDR relationship findings — towards morality.

References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211. [https://doi.org/10.1016/0749-5978\(91\)90020-t](https://doi.org/10.1016/0749-5978(91)90020-t)
- Anglim, J., Morse, G., de Vries, R. E., MacCann, C., & Marty, A. (2017). Comparing job applicants to Non–Applicants using an Item–Level bifactor model on the Hexaco Personality inventory. *European Journal of Personality*, 31(6), 669–684. <https://doi.org/10.1002/per.2120>
- Bäckström, M., & Björklund, F. (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology*, 54(2), 152–159. <https://doi.org/10.1111/sjop.12015>
- Bäckström, M., & Björklund, F. (2020). The properties and utility of less evaluative personality scales: reduction of social desirability; increase of construct and discriminant validity. *Frontiers in Psychology*, 11, 560271. <https://doi.org/10.3389/fpsyg.2020.560271>
- Bäckström, M., & Björklund, F. (2021). Is the General Factor of Personality Really Related to Frequency of Agreeable, Conscientious, Emotionally Stable, Extraverted, and Open Behavior? *Journal of Individual Differences*, 42(3), 148–155. <https://doi.org/10.1027/1614-0001/a000341>
- Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, 43(3), 335–344. <https://doi.org/10.1016/j.jrp.2008.12.013>

- Bäckström, M., Björklund, F., & Larsson, M. R. (2014). Criterion Validity is Maintained When Items Are Evaluatively Neutralized: Evidence from A Full-Scale Five-Factor Model Inventory. *European Journal of Personality*, 28(6), 620–633.
<https://doi.org/10.1002/per.1960>
- Bäckström, M., Björklund, F., Maddux, R. E., & Lindén, M. (2023). The NB5I. *European Journal of Psychological Assessment*, 39(2), 132–140. <https://doi.org/10.1027/1015-5759/a000687>
- Ben-Porath, Y. S., & Tellegen, A. (2020). Minnesota Multiphasic Personality Inventory-3 [Dataset]. In *PsycTESTS Dataset*. <https://doi.org/10.1037/t88933-000>
- Bensch, D., Paulhus, D. L., Stankov, L., & Ziegler, M. (2017). Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment*, 26(3), 351–363.
<https://doi.org/10.1177/1073191117700268>
- Biderman, M. D., Nguyen, N. T., Cunningham, C. J., & Ghorbani, N. (2011). The ubiquity of common method variance: The case of the Big Five. *Journal of Research in Personality*, 45(5), 417–429. <https://doi.org/10.1016/j.jrp.2011.05.001>
- Borkenau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. *European Journal of Personality*, 6(3), 199–214.
<https://doi.org/10.1002/per.2410060303>
- Borkenau, P., & Zaltauskas, K. (2009). Effects of self-enhancement on agreement on personality profiles. *European Journal of Personality*, 23(2), 107–123.
<https://doi.org/10.1002/per.707>

- Bou Malham, P., & Saucier, G. (2016). The conceptual link between social desirability and cultural normativity. *International Journal of Psychology*, 51(6), 474–480.
<https://doi.org/10.1002/ijop.12261>
- Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: effects on criterion-related validity and individual hiring decisions. *Personnel Psychology*, 47(4), 847–860. <https://doi.org/10.1111/j.1744-6570.1994.tb01581.x>
- Connelly, B. S., & Chang, L. (2015). 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>
- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.
- 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>
- De Jonge, P., & Slaets, J. P. (2005). Response sets in self-report data and their associations with personality traits. *The European Journal of Psychiatry*, 19(4).
<https://doi.org/10.4321/s0213-61632005000400002>
- De Vries, R. E. (2010). No evidence for a General Factor of Personality in the HEXACO Personality Inventory. *Journal of Research in Personality*, 45(2), 229–232.
<https://doi.org/10.1016/j.jrp.2010.12.002>

- De Vries, R. E., Hilbig, B. E., Zettler, I., Dunlop, P. D., Holtrop, D., Lee, K., & Ashton, M. C. (2017). Honest people tend to use Less—Not More—Profanity. *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. *Assessment*, 21(3), 286–299. <https://doi.org/10.1177/1073191113504619>
- Dodaj, A. (2012). Social Desirability and Self-Reports: Testing a content and Response-Style model of socially desirable responding. *Europe's Journal of Psychology*, 8(4). <https://doi.org/10.5964/ejop.v8i4.462>
- Dunlop, P. D., Telford, A. D., & Morrison, D. L. (2011). Not too little, but not too much: The perceived desirability of responses to personality items. *Journal of Research in Personality*, 46(1), 8–18. <https://doi.org/10.1016/j.jrp.2011.10.004>
- Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, 86(1), 122–133. <https://doi.org/10.1037/0021-9010.86.1.122>
- Eysenck, S., Eysenck, H., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, 6(1), 21–29. [https://doi.org/10.1016/0191-8869\(85\)90026-1](https://doi.org/10.1016/0191-8869(85)90026-1)
- Feldman, G. (2018). 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>

- Gittelman, S., Lange, V., Cook, W. A., Frede, S. M., Lavrakas, P. J., Pierce, C., Thomas, R. K., Terhanian, G., Bacon, C., & Fulgoni, G. (2015). Accounting for Social-Desirability bias in survey sampling. *Journal of Advertising Research*, 55(3), 242–254. <https://doi.org/10.2501/jar-2015-006>
- Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 11(4), 323–343. <https://doi.org/10.1037/1082-989x.11.4.323>
- Havan, P., & Kohut, M. (2019). Social desirability and personality traits BFI-2. *Acta Psychologica Tyrnaviensia*, 22–23, 5–17
- Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, 39(3), 184–201. <https://doi.org/10.1037/cjbs2007015>
- Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences*, 49(5), 446–450. <https://doi.org/10.1016/j.paid.2010.04.015>
- Holden, R. R., Starzyk, K. B., McLeod, L. D., & Edwards, M. J. (2000). Comparisons among the Holden Psychological Screening Inventory (HPSI), the Brief Symptom Inventory (BSI), and the Balanced Inventory of Desirable Responding (BIDR). *Assessment*, 7(2), 163–175. <https://doi.org/10.1177/107319110000700208>
- Holtrop, D., Hughes, A. W., Dunlop, P. D., Chan, J., & Steedman, G. (2020). Do social desirability scales measure dishonesty? *European Journal of Psychological Assessment*, 37(4), 274–282. <https://doi.org/10.1027/1015-5759/a000607>

- Ingold, P. V., Kleinmann, M., König, C. J., & Melchers, K. G. (2014). Shall we continue or stop disapproving of self-presentation? Evidence on impression management and faking in a selection context and their relation to job performance. *European Journal of Work and Organizational Psychology*, 24(3), 420–432.
<https://doi.org/10.1080/1359432x.2014.915215>
- Konstabel, K., Aavik, T., & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality*, 20(7), 549–566.
<https://doi.org/10.1002/per.593>
- Kwan, V. S. Y., John, O. P., Kenny, D. A., Bond, M. H., & Robins, R. W. (2004). Reconceptualizing individual Differences in Self-Enhancement Bias: an interpersonal approach. *Psychological Review*, 111(1), 94–110. <https://doi.org/10.1037/0033-295x.111.1.94>
- Lee, K., & Ashton, M. C. (2016). Psychometric properties of the HEXACO-100. *Assessment*, 25(5), 543–556. <https://doi.org/10.1177/1073191116659134>
- 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>
- Leite, W. L., & Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne social desirability scale and the balanced inventory of desirable responding. *Educational and Psychological Measurement*, 65(1), 140–154.
<https://doi.org/10.1177/0013164404267285>
- Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, 53, 107–117.
<https://doi.org/10.1016/j.jesp.2014.03.005>

- Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and Self-Deception on the criterion Validity of Personality measures: A Meta-Analysis. *International Journal of Selection and Assessment*, 14(2), 131–141. <https://doi.org/10.1111/j.1468-2389.2006.00339.x>
- Lönnqvist, J., Verkasalo, M., & Bezmenova, I. (2007). Agentic and communal bias in socially desirable responding. *European Journal of Personality*, 21(6), 853–868. <https://doi.org/10.1002/per.639>
- Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the International Personality Item Pool representation of the Revised NEO Personality Inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment*, 26(4), 1070–1084. <https://doi.org/10.1037/pas0000004>
- Marshall, M. B., De Fruyt, F., Rolland, J., & Bagby, R. M. (2005). Socially desirable responding and the factorial stability of the NEO PI-R. *Psychological Assessment*, 17(3), 379–384. <https://doi.org/10.1037/1040-3590.17.3.379>
- McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, 51(6), 882–888. <https://doi.org/10.1037/0022-006x.51.6.882>
- McFarland, L. A., Ryan, A. M., & Kriska, S. D. (2003). Impression management use and effectiveness across assessment methods. *Journal of Management*, 29(5), 641–661. [https://doi.org/10.1016/s0149-2063\(03\)00030-8](https://doi.org/10.1016/s0149-2063(03)00030-8)
- McKelvie, S. J. (2004). Is the Neuroticism Scale of the Eysenck Personality Inventory contaminated by response bias? *Personality and Individual Differences*, 36(4), 743–755. [https://doi.org/10.1016/s0191-8869\(02\)00348-3](https://doi.org/10.1016/s0191-8869(02)00348-3)

- McGee Ng, S.A., Bagby, R. M., Goodwin, B. E., Burchett, D., Sellbom, M., Ayearst, L. E., Dhillon, S., Yiu, S., Ben-Porath, Y. S., & Baker, S. (2015). The effect of response bias on the personality inventory for DSM–5 (PID–5). *Journal of Personality Assessment*, 98(1), 51–61. <https://doi.org/10.1080/00223891.2015.1096791>
- Microsoft Corporation. (2025). *Excel (Microsoft 365 Subscription)*. [Computer Software].
- Microsoft Corporation. <https://www.microsoft.com/en-us/microsoft-365/excel#:~:text=Microsoft%20Excel%20with%20a%20Microsoft,Excel%202007%2C%20and%20Excel%202003>
- Müller, S., & Moshagen, M. (2018). Controlling for response Bias in Self-Ratings of Personality: A comparison of impression management scales and the Overclaiming technique. *Journal of Personality Assessment*, 101(3), 229–236. <https://doi.org/10.1080/00223891.2018.1451870>
- Navarro-González, D., Lorenzo-Seva, U., & Vigil-Colet, A. (2016). How response bias affects the factorial structure of personality self-reports. *Psicothema*, 4(28), 465–470. <https://doi.org/10.7334/psicothema2016.113>
- Nevid, J. S. (1983). Hopelessness, social desirability, and construct validity. *Journal of Consulting and Clinical Psychology*, 51(1), 139–140. <https://doi.org/10.1037/0022-006x.51.1.139>
- 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>
- Ones, D., & Viswesvaran, C. (1998). The Effects of social desirability and faking on Personality and Integrity Assessment for Personnel selection. *Human Performance*, 11(2), 245–269. <https://doi.org/10.1080/08959285.1998.9668033>

Paulhus, D. L. (1991). Balanced Inventory of Desirable Responding--Version 6; Form 40

[Dataset]. In *PsycTESTS Dataset*. <https://doi.org/10.1037/t13624-000>

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., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, 60(2), 307–317. <https://doi.org/10.1037/0022-3514.60.2.307>

Paulhus, D. L., & Trapnell, P. D. (2008). Self-presentation of personality: An agency-communality framework. In O. P. John, R. W. Robins, & L. A. Pervin, *Handbook of personality psychology: Theory and research* (pp. 492-517). Guilford Press.

Pauls, C. A., & Stemmler, G. (2002). Substance and bias in social desirability responding. *Personality and Individual Differences*, 35(2), 263–275. [https://doi.org/10.1016/s0191-8869\(02\)00187-3](https://doi.org/10.1016/s0191-8869(02)00187-3)

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

Peabody, D., & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology*, 57(3), 552–567. <https://doi.org/10.1037/0022-3514.57.3.552>

Pelt, D. H. M., Van Der Linden, D., Dunkel, C. S., & Born, M. P. (2019). The motivation and opportunity for socially desirable responding does not alter the general factor of

personality. *Assessment*, 28(5), 1376–1396.

<https://doi.org/10.1177/1073191119880960>

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, 78(3), 582–593.

<https://doi.org/10.1037/0022-3514.78.3.582>

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, 83(4), 634–644. <https://doi.org/10.1037/0021-9010.83.4.634>

Roth, M., & Herzberg, P. Y. (2007). The resilient type: “Simply the best” or merely an artifact of social desirability? *Psychology Science*, 49(2), 150. http://www.psychologie-aktuell.com/fileadmin/download/PsychologyScience/ps_2_2007_150-167.pdf

Saucier, G. (2002). Orthogonal markers for orthogonal factors: the case of the Big Five. *Journal of Research in Personality*, 36(1), 1–31.

<https://doi.org/10.1006/jrpe.2001.2335>

The jamovi project (2026). jamovi (Version 2.7) [Computer Software]. Retrieved from <https://www.jamovi.org>

Uziel, L. (2010). Rethinking social desirability scales. *Perspectives on Psychological Science*, 5(3), 243–262. <https://doi.org/10.1177/1745691610369465>

Uziel, L. (2013). 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 Der Linden, D., Bakker, A. B., & Serlie, A. W. (2011). The General Factor of Personality in selection and assessment samples. *Personality and Individual Differences*, 51(5), 641–645. <https://doi.org/10.1016/j.paid.2011.05.032>
- Verschuere, B., Uzieblo, K., De Schryver, M., Douma, H., Onraedt, T., & Crombez, G. (2014). The inverse relation between psychopathy and faking good: not response bias, but true variance in psychopathic personality. *Journal of Forensic Psychiatry and Psychology*, 25(6), 705–713. <https://doi.org/10.1080/14789949.2014.952767>
- Vilar, R., Milfont, T. L., & Sibley, C. G. (2020). The role of social desirability responding in the longitudinal relations between intention and behaviour. *Journal of Environmental Psychology*, 70, 101457. <https://doi.org/10.1016/j.jenvp.2020.101457>
- Walker, L. J., & Hennig, K. H. (2004). Differing conceptions of moral exemplarity: just, brave, and caring. *Journal of Personality and Social Psychology*, 86(4), 629–647. <https://doi.org/10.1037/0022-3514.86.4.629>
- Zettler, I., Hilbig, B. E., Moshagen, M., & de Vries, R. E. (2014). 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>

Appendix A

Situations Scale

You will be presented with several real life situations. Please, think of a time in the **past month** when you experienced each of these situations and choose the one option that **BEST** reflects how you responded, and one that reflects your actual response the **LEAST**. It is imperative that you respond **truthfully**.

1. You made a mistake.

- a) I ignored it, b) I accepted the consequences, c) I fixed the mistake as quickly as possible, d) I put the blame on someone else, e) I took responsibility and made sure it wouldn't happen again.

2. A friend let it slip they had a private problem.

- a) I said nothing, b) I let them know I was there for them, c) I helped any way I could, d) I changed the subject, e) I helped and followed up later to see how they were doing.

3. Someone achieved what you wanted for yourself.

- a) I felt discouraged, b) I continued as if nothing happened, c) I worked harder to achieve it too, d) I felt jealous and resented them, e) I congratulated them and used it as motivation.

4. Your advice was rejected.

- a) I got offended, b) I accepted it without further pushing, c) I offered support in their choice, d) I criticized them for not listening, e) I reflected on my advice to improve it for next time.

5. Someone wronged you.

- a) I held a grudge, b) I didn't hold a grudge, c) I made small efforts to improve the relationship, d) I plotted how to get back at them, e) I initiated a calm discussion to clear the air.

6. A friend asked you to hang out when you weren't in the mood.

- a) I said no without elaborating, b) I politely declined with a vague excuse, c) I suggested another time without explaining why I declined, d) I ghosted them, e) I explained honestly and asked to reschedule.

7. You got confused and needed help.

- a) I acted like there wasn't a problem, b) I kept trying to solve the issue myself, c) I reached out and sought help, d) I blamed others for not making things clearer, e) I asked for clarification so I wouldn't get confused again.

8. Something made you feel stressed.

- a) I did nothing, b) I distracted myself by positive things, c) I tried to resolve the root of stress, d) I snapped at others, e) I managed my stress and reflected on how to prevent it in the future.

9. You got bored and felt like doing something different.

- a) I distracted myself temporarily, b) I decided against it and powered through, c) I tried reframing it to feel more manageable, d) I zoned out and lost track of what I was doing, e) I managed to turn it into something interesting.

10. You found yourself in a heated group argument.

- a) I was unreceptive to others' points, b) I didn't engage in the argument, c) I calmly participated, d) I took sides and escalated the conflict, e) I tried to mediate and help everyone be heard.

11. You encountered a hurdle in your work.

- a) I kept doing the same thing even though it wasn't working, b) I managed to work around it, c) I came up with a new, creative solution, d) I gave up entirely, e) I sought feedback and used the obstacle to improve myself.

12. Something you worked hard on wasn't working out, even though you followed instructions.

- a) I got discouraged, b) I took a break and returned to it later, c) I asked for help, d) I got angry and gave up, e) I reflected on the failure and tried a new strategy.

13. You got into an argument with a person close to you and it made you feel bad afterwards.

- a) I pretended nothing happened, b) I showed I was ready to move on with my gestures, c) I apologised in hopes of reconciliation, d) I ignored them and acted coldly, e) I took responsibility and asked them if we could talk it through.

14. You were in pain.

- a) I let it affect my mood, b) I quietly dealt with it so it wouldn't bother others, c) I did everything I could to stop the pain, d) I exaggerated it and complained constantly, e) I reflected on what might have caused it so I could prevent it in the future.

15. Your friend was late for a meet-up and that made you upset.

- a) I was put off once they arrived, b) I thought they probably had a reason and forgot about it quickly, c) I voiced my feelings calmly, d) I made a comment to make them feel bad, e) "I politely told them how I felt and asked to plan better in the future.

16. There was a situation in which you thought you were right, but the person in authority disagreed.

a) I fought back, b) I accepted the decision of authority, c) I tried to convince them with logical arguments, d) I dismissed them and stuck to my view, e) I sought to understand their view and find common ground.

17. A stranger was rude to you.

a) I was rude back, b) I ignored it, c) I calmly confronted them about it, d) I directly insulted them, e) I was assertive and attempted to change their behaviour constructively.

18. You disliked a person in a group of friends you were in.

a) I gossiped about them with some of the other friends, b) I just tried my best to avoid them and focused more on others, c) I attempted to get over my feelings and make friends with them, d) I tried to get others to dislike them too, e) I initiated a positive interaction to reset our dynamic.

19. You were talking with a person who you thought was quite boring.

a) I cut the conversation short abruptly, b) I politely listened to them, c) I tried my best to actively engage in conversation nonetheless, d) I was not subtle with my body language (e.g., I rolled my eyes, looked “through” them...), e) I did my best to make the conversation more interesting.

20. You heard of a really bad thing happening to a person you once knew, but are not close to anymore.

a) The news did not bother me, b) I felt bad for them but it wasn't any of my concern, c) I reached out to see how they were, d) I joked about it with others, e) I offered them support.

Appendix B

Honesty Scale

Below, you will find items relating to your attitudes, norms and control regarding honesty. Your task is to choose a number on the scale from **1 (highly disagree) to 5 (highly agree)**, indicating your agreement with the following statements pertaining to your experience. Try your best to answer **each question** and to answer it **truthfully**.

Attitudes

1. I believe honesty is one of the most important traits a person can have.
2. In my relationships, honesty is the trait I value the most.
3. It is better to be hurt by the truth than sheltered by a lie.
4. People who deceive others are being disrespectful.
5. Sometimes, lying is acceptable to achieve a goal. (*reverse-coded*)
6. Lying can sometimes be the best option. (*reverse-coded*)
7. If everyone were honest, the world would be a better place.

Norms

1. Because people around me are honest, they expect me to be honest too.
2. If I lie, others would frown upon it.
3. I feel socially obligated to be sincere around my friends.
4. Most people I know expect others to be honest.
5. My parents expect me to be honest.

6. People around me would criticise me if I didn't tell the truth.
7. Being honest is generally expected in my environment.

Perceived behavioural control:

1. Telling the truth or lying is completely under my control.
2. Even when it's hard, I can choose to be honest.
3. No one can stop me from lying if I want to.
4. I can avoid situations where I'd feel pressured to lie.
5. I am usually in control of how much of the truth I can share.
6. Sometimes, circumstances make it hard for me to be honest. (*reverse-coded*)
7. It's not just up to me whether to be honest or lie. (*reverse-coded*)