Measurement Instruments
for the Social Sciences

**INTERNATIONAL ADAPTATION OF
MEASUREMENT INSTRUMENTS**

Open Access

CrossMark

An English-Language Adaptation of the Social Desirability–Gamma Short Scale (KSE-G)

Désirée Nießen[1*] , Melanie V. Partsch[1], Christoph J. Kemper[2] and Beatrice Rammstedt[1]

Abstract

The Social Desirability–Gamma Short Scale—the English-language adaptation of the Kurzskala Soziale Erwünschtheit–Gamma (KSE-G)—measures two aspects of the Gamma factor of socially desirable responding (SDR): exaggerating positive qualities (PQ+) and minimizing negative qualities (NQ–). The items of the German-language source version were translated into English using the TRAPD approach. Our empirical validation shows that the reliability and validity coefficients of the English adaptation are comparable to those of the German source instrument. Moreover, the results of measurement invariance testing suggest metric measurement invariance of the scale for the United Kingdom and Germany, thus implying comparability of correlations based on the latent factors across the two countries.

Keywords: Social desirability, Impression management, Gamma, Short scale, Response behavior

Introduction

Social desirability is a response tendency that biases individual item responses, thereby leading to deviation from true scores. High scores on socially desirable responding (SDR) scales, and high correlations between SDR scales and self-report instruments, indicate a possible distortion of respondents' answers on self-report questionnaires (Paulhus & Trapnell, 2008). There are a number of different theoretical traditions of the conceptualization of SDR. One of the most recent of these conceptualizations is bi-dimensional, with the two factors Gamma and Alpha. As opposed to Alpha, scales assessing the Gamma factor are particularly suitable for checking whether self-report questionnaire responses measuring behavior, personality, and attitudes are biased by SDR. Several scales assessing both SDR factors and/or focusing on individual diagnostics already exist. However, these scales are comparatively lengthy. In order to provide a measure of the Gamma factor of SDR that can be used also for research purposes with extreme time limitations, Kemper, Beierlein, Bensch, Kovaleva, and Rammstedt (2014) developed for the

German context the KSE-G (Kurzskala Soziale Erwünschtheit–Gamma [Social Desirability–Gamma Short Scale]). Due to its short completion time (< 1 min), the instrument can be applied in research settings with severe time limitations, for example, large-scale surveys, and can be used to check whether questionnaire responses are biased by SDR. The German-language KSE-G has been validated for the adult population in Germany, irrespective of age and social class. To enhance its usability, the authors of the scale translated and adapted the items to English. However, an empirical investigation of the appropriateness of this adaptation was hitherto lacking. Such validation is the only way to test the applicability of the English KSE-G to an English-speaking population. The aim of the present study, therefore, was to conduct a comprehensive validation study of the English adaptation of the KSE-G and to compare its psychometric properties directly with those of the German source version.

Theoretical background

Socially desirable responding is defined as the "tendency to give overly positive self-descriptions" (Paulhus, 2002, p. 50) "in order to put forward a more socially acceptable self-image" (Haghighat, 2007). The construct has been investigated in psychological research for over 60

* Correspondence: desiree.niessen@gesis.org
[1]GESIS – Leibniz Institute for the Social Sciences, P.O. Box 12 21 55, 68072 Mannheim, Germany
Full list of author information is available at the end of the article

years now. There is a broad range of approaches operationalizing SDR, and many scales have been developed over the years (for an overview, see, e.g., Paulhus, 1991a; Paulhus, 2002; Paulhus & Trapnell, 2008).

A widespread, comprehensive, and integrative conceptualization of the SDR construct was developed over the years by Delroy Paulhus. Initially, Paulhus (1984, 1986) assumed that SDR consisted of the two relatively independent factors: (conscious) *impression management* (IM)—also known as *Gamma*—and (unconscious) *self-deceptive enhancement* (SDE)—also known as *Alpha* (Wiggins, 1964).[1] However, a series of studies with instructional manipulations yielded evidence for associating Gamma and Alpha with the so-called Big Two (Paulhus & Trapnell, 2008), namely *communion* and *agency* (Bakan, 1966). This research indicated that respondents interpreted instructions to "respond in a socially desirable way" to mean that they should claim communal attributes (e.g., responsibility, cooperativeness), which led to higher scores on Gamma measures than on Alpha measures (Paulhus & Trapnell, 2008, p. 502). By contrast, respondents interpreted instructions to "respond as if you are strong and competent" to mean that they should claim agentic attributes (i.e., prominence, status; Paulhus & John, 1998), which resulted in higher scores on Alpha measures than on Gamma measures (Paulhus, Tanchuk, & Wehr, 1999, as cited in Paulhus & Trapnell, 2008, p. 502).

Based on these findings, Paulhus (2002) developed an integrative model (further elucidated in Paulhus & Trapnell, 2008), in which he considered both a content distinction of SDR (communion- vs. agency-induced SDR) and an audience distinction (IM induced by a public audience vs. SDE induced by a private audience, i.e., the self). From this integrative model, a revised Gamma factor (communion) of SDR and a revised Alpha factor (agency) were derived, both of which have IM and SDE components.

Communion-related SDR (i.e., the revised Gamma factor) involves "excessive adherence to group norms and minimization of social deviance" (Paulhus & Trapnell, 2008, p. 498); it is related to qualities such as cooperativeness, warmth, and dutifulness. *Communion management* describes the communal aspect of IM and "involves excuse making and damage control" (Paulhus & Trapnell, 2008, p. 503). *Moralistic bias* describes the communal aspect of SDE. It is defined as a "self-deceptive tendency to deny socially deviant impulses and to claim sanctimonious 'saint-like' attributes" (Paulhus & John, 1998, p. 1026). This tendency manifests itself in "overly positive self-perceptions" on personality traits associated with communion, such as "agreeableness, dutifulness, and restraint" (Paulhus & John, 1998, p. 1026). In contrast, agency-related SDR (i.e., the revised Alpha factor) involves "exaggerated achievement striving and

self-importance" (Paulhus & Trapnell, 2008, p. 498) and is associated with qualities such as strength, competence, and cleverness. *Agency management* describes the agentic aspect of IM and manifests itself, for example, in bragging. *Egoistic bias* describes the agentic aspect of SDE and is understood as a self-deceptive tendency to exaggerate one's social and intellectual status. This leads to unrealistically positive self-perceptions on personality traits associated with agency, such as "dominance, fearlessness, emotional stability, intellect, and creativity" (Paulhus & John, 1998, p. 1026).

The two SDR factors are associated with different personality traits. Gamma shows the strongest positive correlations with Agreeableness, followed by Conscientiousness and, to a lesser extent, Emotional Stability. Alpha shows the strongest positive correlations with Emotional Stability, followed by Conscientiousness, Extraversion, and, to a lesser extent, Openness and Agreeableness (Hart, Ritchie, Hepper, & Gebauer 2015; Li & Bagger, 2006; Paulhus, 1988).

In the early years of SDR research, scales were not designed to distinguish between different facets of SDR but rather were constituted as unidimensional measures linked to different conceptions of SDR (e.g., the Edwards Social Desirability Scale [ESD], Edwards, 1957; the Wiggins Social Desirability Scale [Wsd], Wiggins, 1959; the Marlowe-Crowne Social Desirability Scale [MC-SDS], Crowne & Marlowe, 1960). In contrast, Paulhus (1991b, 1998) developed a two-dimensional measure—the Balanced Inventory of Desirable Responding (BIDR)—based on his concept of the two dimensions of SDR, namely IM and SDE. This approach is widely accepted in the current research on SDR (e.g., Asgeirsdottir, Vésteinsdóttir, & Thorsdottir, 2016; Hart et al. 2015; Stöber 2001; Wiggins, 2003). Short measures were derived from the full scales measuring either a unidimensional or a bi-dimensional SDR concept. These short scales include the Balanced Inventory of Desirable Responding Short Form (BIDR-16; Hart et al., 2015), which comprises 16 items, or the Social Desirability Scale-17 (SDS-17; Stöber, 2001), which consists of 17 items.

However, as numerous studies (e.g., Paulhus, 1984; Paulhus & Reid, 1991; but see Li & Bagger, 2006) had indicated that Gamma seemed to bias self-reported behavior, personality characteristics, and attitudes more than Alpha, a scale was needed that identified a person's tendency for socially desirable responding in terms of Gamma. Social-scientific self-report surveys often refer to the social significance of the survey in order to increase the willingness to participate. In such situations, the moralistic bias, which is induced by the assessment setting, could be increased further. Respondents could therefore strive to answer like a "nice person," "well socialized," or "good person" (Paulhus & Trapnell, 2008)

leading to deviation from true scores. In particular, an ultra-short instrument was lacking that was suitable even for extremely time-restricted surveys and that tapped only the relevant comprehensive and revised understanding of Gamma encompassing communion management and moralistic bias.

That is why Kemper et al. (2014) developed the KSE-G, a short scale to assess the Gamma SDR factor, that is, communion-induced SDR reflected both in IM and SDE. When constructing the scale, they identified two subscales of SDR–Gamma. To be more precise, Kemper et al. (2014) did not expect to find these two dimensions. Instead, they detected them factor analytically, checked them with a confirmatory factor analysis (CFA), and were able to replicate them in the further construction process. Following Roth, Snyder, and Pace (1986), who also found these two dimensions, they labeled these subscales *exaggerating positive qualities* and *minimizing negative qualities* of the self.[2] Subscales were considered to be somewhat related but largely independent internally homogenous item clusters which reflect that some respondents "systematically overreport their performance of a wide variety of desirable behaviors and underreport undesirable behaviors" (Paulhus, 1991a, p. 37). The items of one dimension describe polite, sociable, and adapted behaviors that are socially desirable but rare, and the items of the other one describe inappropriate behaviors that are socially undesirable but frequent. These contents are intended to reflect Gamma values (communion) in particular.

Scale development

To develop the KSE-G, Kemper et al. (2014) drew on items from existing social desirability scales, such as the Soziale-Erwünschtheits-Skala-17 (SES-17; Stöber, 1999), a German-language adaptation of the SDS-17 (Stöber, 2001) and a German-language adaptation of the MC-SDS (Lück & Timaeus, 1969). These items were revised to make them more comprehensible and content valid. The revised items were then tested using item and structural analysis. In an iterative process, the authors discarded some items and replaced them with newly developed ones (for more detailed information, see Kemper et al., 2014). The German-language KSE-G was thoroughly validated based on a comprehensive sample that reflected the adult German population. To enhance the usability of the KSE-G, the scale was translated and adapted to English by translating the items following the so-called TRAPD approach (Harkness, 2003). First, two professional translators (native speakers) translated the items independently of each other into British English and American English, respectively. Second, an alignment meeting was held where psychological experts, the two translators, and an expert in questionnaire

translation reviewed the various translation proposals and developed the final translation.

The source instrument by Kemper et al. (2014) was developed in and validated for the German language. The aim of the present study was to validate the English-language adaptation of the KSE-G and to directly compare its psychometric properties with those of the German source version. In line with earlier findings, we expected strongest correlations with Agreeableness, followed by Conscientiousness and Emotional Stability, and small correlations with Openness and Extraversion (Hart et al., 2015; Kemper et al., 2014; Li & Bagger, 2006; Paulhus, 1988; Paulhus, 2002; Stöber, 2001).

Method

Samples

To investigate the psychometric properties of the English adaptation of the KSE-G, and their comparability with those of the German source instrument, we assessed both versions in a web-based survey (computer-assisted self-administered interviewing [CASI]) conducted in the United Kingdom (UK) and in Germany (DE) by the online access panel provider respondi AG. Fielding took place in January 2018. For both countries, quota samples were drawn that reflected the heterogeneity of the adult population with regard to age, gender, and educational attainment. Only native speakers of the respective languages were recruited. Respondents were financially rewarded for their participation. In both countries, a subsample was reassessed after approximately three to four weeks ($Mdn_{UK} = 28$ days; $Mdn_{DE} = 20$ days).

Only respondents who completed the full questionnaire—that is, who did not abort the survey prematurely—were included in our analyses. To handle missing values on single items, we used full information maximum likelihood estimation (FIML) in our analyses. This yielded gross samples of $N_{UK} = 508$ and $N_{DE} = 513$, respectively. In the next step, invalid cases were excluded based on (a) ipsatized variance, that is, the within-person variance across items (Kemper & Menold, 2014), if the person fell within the lower 5% of the sample distribution of ipsatized variance; (b) the Mahalanobis distance of a person's response vector from the average sample response vector (Meade & Craig, 2012) if he/she fell within the upper 2.5% of the sample distribution of the Mahalanobis distance; and (c) response time if the person took, on average, less than 1 s to respond to an item. Our intention in choosing relatively liberal cutoff values was to avoid accidentally excluding valid cases and thereby creating a systematic bias in our data. The outlined approach resulted in total exclusion of 7.9% of cases in the UK subsample and 7.6% of cases in the DE subsample, yielding net sample sizes of $N_{UK} = 468$ (retest: $N_{UK} = 111$) and $N_{DE} = 474$ (retest:

N_{DE} = 117), respectively. Table 1 depicts in detail the sample characteristic features and distribution.

Material
The online survey was conducted in German for the German sample and in English for the UK sample. It comprised the respective language versions of the KSE-G.

The KSE-G consists of six items covering the two aspects of the Gamma factor of social desirability, namely exaggerating positive qualities (PQ+) and minimizing negative qualities (NQ–). The English adaptations of these items are displayed in Table 2 and in the Additional file 1 in the Supplementary Online Material (for the original German items, see Additional file 2 in the Supplementary Online Material and Kemper et al., 2014). As in the German source instrument, all items are formulated positively in the direction of the underlying aspect. Items are answered using a 5-point rating scale ranging from *doesn't apply at all* (1) to *applies completely* (5).[3] The scale score of social desirability is computed separately for each subscale (PQ+ and NQ–). For this purpose, the unweighted mean score of the three items of each subscale is computed.[4]

In addition to administering the KSE-G, a set of sociodemographic variables (gender, age, highest level of education, income, and employment status) was assessed.

To validate the KSE-G against the Big Five dimensions of personality, a short scale measure of the Big Five, the extra-short form of the Big Five Inventory–2 (BFI-2-XS; English version: Soto & John, 2017; German version: Rammstedt, Danner, Soto, & John, 2018), was also administered as part of the survey.[5]

Results
To validate the English adaptation of the KSE-G, and to investigate its comparability with the German source version, we analyzed psychometric criteria—more precisely, reliability and validity—in both language versions. Moreover, we assessed test fairness across both countries via measurement invariance tests. The statistical analysis

was run with R; the code can be found in the Additional file 3 in the Supplemetary Online Material.

Descriptives and reference ranges
In the first step, we report the descriptive statistics and reference ranges separately for both versions of the KSE-G. Table 3 shows the means, standard deviations, skewness, and kurtosis for the six items, as well as reliability coefficients for both subscales of the KSE-G separately for the English and German samples. Additional file 4: Table S1 in the Supplementary Online Material indicates the reference ranges in terms of means, standard deviations, skewness, and kurtosis of the two subscales of the KSE-G for the total population, as well as separately for gender and age groups.

Reliability
As estimates for the reliability of the KSE-G, we computed Cronbach's alpha (Cronbach, 1951), McDonald's omega (McDonald, 1999; Raykov, 1997), and the test-retest stability for the two subscales PQ+ and NQ–. The rationale for using these measures was twofold. First, we wanted to provide information on the most commonly used reliability estimate, namely Cronbach's alpha, even though the appropriateness of this measure of internal consistency is limited in the case of ultra-short scales, in which items are selected to reflect the bandwidth of the underlying dimension (i.e., its heterogeneity but not its homogeneity). Second, we report McDonald's omega, as a more appropriate measure in the current context, because we specified a tau-congeneric model, and each subscale consists of only three items.

The reliability estimates (see Table 3) ranged between .65 and .67 (UK) and between .70 and .72 (DE) for PQ+ and between .64 and .79 (UK) and between .67 and .69 (DE) for NQ–, which can be deemed sufficient for research purposes (Aiken & Groth-Marnat, 2006; Kemper, Trapp, Kathmann, Samuel, & Ziegler, 2018). In detail, PQ+ proved to be more reliable in Germany than in the UK, whereas NQ– showed even better reliability estimates in the UK than in Germany

Table 1 Sample characteristic features

	United Kingdom	Germany
N	468	474
Mean age in years (SD) [range]	45.2 (14.5) [18–69]	44.0 (14.4) [18–69]
Proportion of women (%)	52.6	50.0
Educational level (%)		
Low: never went to school, skills for life/1–4 GCSEs A*–C or equivalent	34.8	33.5
Middle: 5 or more GCSEs A*–C/vocational GCSE/GNVQ intermediate or equivalent	32.1	33.8
High: 2 or more A-levels or equivalent	33.1	32.7

Note. The equivalent German educational levels were as follows (from low to high): *ohne Bildungsabschluss/Hauptschule* [no educational qualification; lower secondary leaving certificate], *mittlerer Schulabschluss* [intermediate school leaving certificate], *(Fach-)Hochschulreife* [higher education entrance qualification]

Table 2 Items of the English-Language Adaptation of the Social Desirability–Gamma Short Scale

No.	Item	Subscale
1	In an argument, I always remain objective and stick to the facts.	PQ+
2	Even if I am feeling stressed, I am always friendly and polite to others.	PQ+
3	When talking to someone, I always listen carefully to what the other person says.	PQ+
4	It has happened that I have taken advantage of someone in the past.	NQ–
5	I have occasionally thrown litter away in the countryside or on to the road.	NQ–
6	Sometimes I only help people if I expect to get something in return.	NQ–

Note. The instructions are as follows: "The following statements may apply more or less to you personally. Please indicate to what extent they apply to you."
PQ+ = exaggerating positive qualities; NQ– = minimizing negative qualities

(except in the case of test-retest stability). As internal consistency estimates vary across groups, test-retest correlations are recommended for a comparison of the reliability of scale scores.

Validity

Besides content-related validity, which was ensured by Kemper et al. (2014) within the original scale development process, we investigated two types of validity: factorial validity and construct validity. Content-related validity "refers to the degree to which the test content elicits behaviors that are representative of the universe of construct-related behaviors the test is designed to measure" (Kemper, 2017, p. 1). Factorial validity is "the validity of a test determined by its correlation with a factor [...] determined by factor analysis" (Colman, 2009). Construct validity is "the degree to which a test measures what it claims, or purports, to be measuring" (Brown, 1996, p. 231).

We first investigated the factorial structure of the KSE-G in the UK and DE in two separate CFAs. As the fit indices proved to be acceptable to good,[6] we subsequently conducted multi-group confirmatory factor analysis (MG-CFA) using a two-dimensional measurement model developed for Germany by Kemper et al. (2014) with two intercorrelated latent factors capturing PQ+ and NQ–. In both countries, factor loadings and item
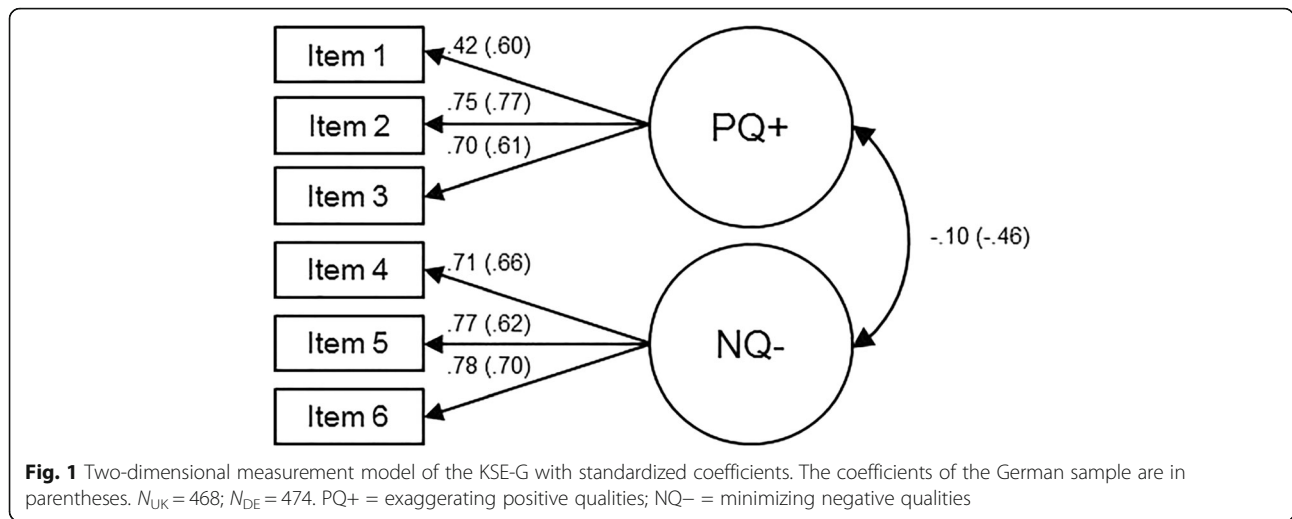
intercepts were freely estimated, whereas the variance of the latent PQ+ and NQ– factor was set to 1. We used robust maximum likelihood estimation (MLR). The model is plotted in Fig. 1; its fit indices suggest an acceptable to good model fit (Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger, & Müller, 2003; Schweizer, 2010). The fit indices refer to the commonly used MLR-scaled RMSEA and CFI indices, which—strictly speaking—only apply to populations: $\chi^2(16) = 58.032$ (UK: $\chi^2 = 33.014$; DE: $\chi^2 = 25.017$), $p < .001$, CFI = .956, RMSEA = .075, SRMR = .049.[7] The size of the items' factor loadings confirms the two-dimensional measurement model, too (see Fig. 1), and gives a first indication of the factorial validity of the scale.

Convergent and discriminant construct validity was computed based on manifest correlations. The correlation coefficients are depicted in Table 4; their interpretation is based on Cohen (1992): small effect ($r \geq .10$), medium effect ($r \geq .30$), and strong effect ($r \geq .50$). Due to alpha accumulation through multiple testing, only coefficients with a significance level above $p < .001$ are interpreted (this is the threshold after Bonferroni adjustment—we use adjusted significance levels only to decide which significant correlations should be used for interpretation; Table 4 displays unadjusted p-values). Before computing the correlations, we recoded the items of NQ–. Hence, high scores on PQ+ are tantamount to high scores

Table 3 Descriptive statistics for KSE-G items and subscales

Item	M		SD		Skewness		Kurtosis		α		Ω		r_{tt}	
	UK	DE	UK	DE	UK	DE	UK	DE	UK	DE	UK	DE	UK	DE
PQ+	3.55	3.67	0.73	0.68	− 0.23	− 0.20	− 0.13	− 0.36	.65	.70	.67	.70	.64	.72
1	3.40	3.46	0.93	0.89	− 0.28	− 0.31	− 0.47	− 0.32						
2	3.47	3.57	1.05	0.89	− 0.42	− 0.20	− 0.52	− 0.54						
3	3.77	3.96	0.88	0.81	− 0.37	− 0.54	− 0.45	0.04						
NQ–	1.90	2.12	0.93	0.89	1.29	0.87	1.07	0.48	.79	.69	.79	.69	.64	.67
4	2.25	2.30	1.19	1.09	0.75	0.56	− 0.32	− 0.46						
5	1.76	2.11	1.10	1.25	1.44	0.85	1.22	− 0.48						
6	1.69	1.94	1.04	1.04	1.46	1.03	1.23	0.47						

Note. UK = United Kingdom (*N* = 468; retest: *N* = 111); DE = Germany (*N* = 474; retest: *N* = 117); PQ+ = exaggerating positive qualities; NQ– = minimizing negative qualities. The time interval between test and retest ranged between three to four weeks (Mdn_{UK} = 28 days; Mdn_{DE} = 20 days)

Fig. 1 Two-dimensional measurement model of the KSE-G with standardized coefficients. The coefficients of the German sample are in parentheses. $N_{UK} = 468$; $N_{DE} = 474$. PQ+ = exaggerating positive qualities; NQ– = minimizing negative qualities

on NQ–, implying high SDR. In order to investigate both types of construct validity by examining whether an underlying moralistic bias in answering personality items existed, we correlated the two subscales of the KSE-G with the Big Five traits, Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness, assessed with the BFI-2-XS (Rammstedt et al., 2018; Soto & John, 2017). The results (see Table 4) support our expectations: For both countries, and for both subdimensions, the strongest associations were found for Agreeableness, followed by Conscientiousness. Stable across the two countries, we found also substantial associations of PQ+ with Emotional Stability and Openness. Small or zero effects were found for Extraversion. In sum, the pattern of correlations

confirms construct validity and points toward a moralistic bias in the respondents' answers.

Furthermore, we calculated correlations between the two Gamma factors of the KSE-G and relevant sociodemographic variables, namely employment status, income, educational level, age, and gender. Only a little evidence exists to date on sociodemographic and socioeconomic correlates of SDR. In their initial validation study of the German KSE-G, Kemper et al. (2014) reported a small positive association between age and PQ+, a medium positive association between age and NQ–, and a small positive association between gender and NQ–. The present analyses partly support these associations both for the German source version and its English adaptation. There were small to medium correlations between NQ– and employment status (UK only), age, and gender. Individuals with a high employment status, and elderly individuals, had a greater tendency to minimize negative qualities. Men were less likely to minimize negative qualities than women. There were no associations between educational level and either PQ+ or NQ– and no reportable associations between all sociodemographic variables and PQ+.

Table 4 Correlations of the KSE-G with relevant variables

	PQ+		NQ–	
	UK	DE	UK	DE
Big Five				
Extraversion	.13**	.04	− .13**	− .03
Agreeableness	**.38***	**.41***	**.31***	**.35***
Conscientiousness	**.20***	**.37***	**.29***	**.26***
Emotional Stability	**.30***	**.31***	.08	.12**
Openness	**.21***	**.25***	− .01	.09
Employment status	.11*	.04	**.20***	.10*
Income	.05	.07	− .16**	.01
Educational level	.04	.05	− .06	− .07
Age	.09*	.11*	**.27***	**.33**
Gender	− .04	.05	**.21***	**.18***

Note. UK = United Kingdom; $N = 468$; $N_{Employment\ status} = 450$; $N_{Income} = 431$); DE = Germany ($N = 474$; $N_{Employment\ status} = 462$; $N_{Income} = 449$); PQ+ = exaggerating positive qualities; NQ– = minimizing negative qualities. Gender: 1 = male, 2 = female. *$p < .05$, **$p < .01$, ***$p < .001$. Coefficients significant on the $p < .001$ level are set in boldface

International equivalence and fairness
We assessed test fairness across countries via measurement invariance tests with MG-CFA (Vandenberg & Lance, 2000; Widaman & Reise, 1997). In order to determine the level of measurement invariance, we used the cutoff values recommended by Chen (2007). According to these benchmarks, SRMR as well as MLR-scaled CFI and RMSEA indicate metric measurement invariance of the two subscales across the United Kingdom and Germany, implying comparability of correlations based on the latent factors between both countries (configural

model: CFI = .956, RMSEA = .075, SRMR = .049; metric model: CFI = .951, RMSEA = .071, SRMR = .052; scalar model: CFI = .935, RMSEA = .074, SRMR = .056).[8]

Discussion and conclusion

The aim of the present study was to validate the English-language adaptation of the Social Desirability–Gamma Short Scale (KSE-G; Kemper et al., 2014), an ultra-short scale assessing the Gamma factor of SDR. The scale was constructed for use in assessment settings with severe time limitations, such as large-scale surveys. In survey conditions, communal behavior—and thus a moralistic bias in respondents' answers—may be evoked. The KSE-G was developed to detect this bias. Our results—based on two comprehensive samples representing the heterogeneity of the UK and German adult populations—reveal, first, that the psychometric properties of the English adaptation of the KSE-G are comparable to those of the German source version. Second, our findings indicate that the English version of the KSE-G is also a valid and useful instrument for detecting socially desirable responding tendencies in research settings with extreme time limitations.

In detail, we were able to replicate the two-dimensional structure of the Gamma factor of SDR that Kemper et al. (2014) conducted when constructing the KSE-G. In addition, the estimates for reliability indicate acceptable scale scores for the English adaptation compared to the German source version. Furthermore, the results of measurement invariance testing suggest metric measurement invariance of the scale, thereby implying comparability of correlations based on the latent factors across countries. As measurement invariance testing could not confirm scalar invariance, it would be necessary to test the comparability of the KSE-G scale scores across gender and age groups more closely. In our study, sample sizes are too small for subgroup comparisons, but future research should have a deeper look at it.

Also with regard to the scale's construct validity, we could partly support the findings for the German source version: Like Kemper et al. (2014), we found the strongest correlations with Agreeableness and Conscientiousness and the smallest/zero correlations with Extraversion for both subscales and countries. Individuals who were high in Big Five Agreeableness and Conscientiousness had a tendency to exaggerate positive qualities and to minimize negative qualities. However, unlike Kemper et al. (2014), who found small associations of Emotional Stability and Openness with both subscales, we found substantial and strong associations for both countries, but only for the PQ+ subscale. Individuals who were emotionally stable or open were prone to exaggerate positive qualities. This highlights the need to have a closer look at the two subscales separately, an essential aspect that extends the work

of Kemper et al. (2014). As past studies have found the strongest correlations between Agreeableness and Conscientiousness and IM (e.g., Hart et al., 2015; Li & Bagger, 2006; Paulhus, 2002; Stöber, 2001), NQ– seems to depict the IM component of Gamma. In contrast, in past studies, Emotional Stability has been found to be the strongest correlate of SDE, followed by Conscientiousness, Extraversion, Agreeableness, and Openness (Hart et al., 2015; Li & Bagger, 2006; Paulhus, 1991a). Evidence reported by Paulhus (2002) suggests that SDE may even play a role in all personality dimensions. Although the relations between PQ+ and Extraversion were negligible, the results allow us to conclude that PQ+ seems to depict the SDE component of Gamma.

Results of the descriptives and the factor loadings also point towards a content distinction of the two subdimensions. The intercorrelation between PQ+ and NQ– is quite small in the UK indicating two distinct and mostly independent subdimensions. Moreover, although still reasonable, it is apparent that NQ– is more right-skewed than PQ+, particularly for the UK. One possible reason might be the abovementioned different contents of the subdimensions. PQ+ is associated with SDE, whereas NQ– is associated with IM and therefore even more susceptible to SDR. Our study provides the first attempt to distinguish between the two subscales in terms of content in more detail. In order to gain an even deeper understanding of the different contents and concepts of PQ+ and NQ–, future research is needed. In addition, although there are enough indications for construct validity for the German source version of the KSE-G, a more comprehensive validation of the English version (with scales of similar SDR constructs, of constructs that are related but conceptually distinct from SDR, and of constructs that distinguish the two subscales of the KSE-G) would certainly be desirable.

The scope of our study was limited in several ways. First, factor correlations across countries are quite different. At this point, no decision can be made whether it is due to content-related culturally different proximity of the two subscales or due to the language adaptation. Second, our samples were restricted to participants in a web-based survey (CASI). Hence, we cannot generalize our findings to the population as a whole, including, for example, non-computer-literate persons. Furthermore, we were unable to investigate the psychometric properties, and especially the scale means, for different assessment modes, in particular, interviewer-based modes. As face-to-face or telephone interviewing situations, for example, have been found to encourage SDR (e.g., Bowling, 2005; Duffy, Smith, Terhanian, & Bremer, 2005; Holbrook, Green, & Krosnick, 2003; Kaminska & Foulsham, 2013) by evoking, in particular, communal behavior, it is possible

that higher SDR scores, on average, might be found in such modes. Finally, our validation of the English-language KSE-G was restricted to the population of the UK only. As a consequence, the results are not automatically generalizable to other English-speaking populations, for example, in the United States. Future studies should address these limitations.

In sum, the results of the present study show for the first time the utility of the English-language adaptation of the KSE-G and the comparability of its psychometric properties with those of the German source version. Researchers in English-speaking countries now have the possibility to assess the Gamma factor of SDR in settings with severe time limitations in order to investigate whether questionnaire responses are (moralistically) biased by SDR leading to deviation from true scores. It is recommended to use the scale in social-scientific self-report surveys—especially when measuring behavior, personality characteristics, and attitudes.

Endnotes

[1]The neutral designations Gamma and Alpha are two of a total of six for two out of six factors representing stylistic scales in the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1951).

[2]The exact labels in Roth et al. (1986) are *attributing the existence of positive characteristics* and *denying the presence of negative qualities.*

[3]Note that Kemper et al. (2014) coded the items from 0 to 4.

[4]We suggest that individual answers should be aggregated to the scale level only if there are no missing values.

[5]As the KSE-G was administered as part of a comprehensive online survey for the validation of various scales, there was no room for further validation scales of similar SDR constructs, of constructs that are related but conceptually distinct from SDR, and of constructs that distinguish the two subscales of the KSE-G. However, there are already enough indications of construct validity for the German source scale (see Kemper et al., 2014).

[6]UK—$\chi^2(8)$ = 25.786, *p* < .01, CFI = .965, RMSEA = .069, SRMR = .050; DE—$\chi^2(8)$ = 32.059, *p* < .001, CFI = .947, RMSEA = .080, SRMR = .048.

[7]Taking the sample size into account prevents biased fit indices, yielding so-called robust CFI and robust RMSEA values in R/lavaan (Brosseau-Liard, Savalei, & Li, 2012; Brosseau-Liard & Savalei, 2014): robust CFI = .963, robust RMSEA = .077.

[8]Robust CFI and robust RMSEA are as follows: configural model—robust CFI = .963, robust RMSEA = .077; metric model—robust CFI = .958, robust RMSEA = .074; scalar model—robust CFI = .945, robust RMSEA = .077.

Additional files

Additional file 1: Answer Sheet (English Version) KSE-G. (PDF 155kb)

Additional file 2: Answer Sheet (German Version) KSE-G. (PDF 155kb)

Additional file 3: R Code for Analysis. (PDF 195kb)

Additional file 4: Table S1. Reference Ranges of the KSE-G for the Total Population and for Gender and Age Groups. (PDF 171kb)

Abbreviations

BFI-2-XS: Big Five Inventory–2 Extra-Short Form; BIDR: Balanced Inventory of Desirable Responding; BIDR-16: Balanced Inventory of Desirable Responding Short Form; CASI: Computer-assisted self-administered interviewing; CFA: Confirmatory factor analysis; DE: Germany; ESD: Edwards Social Desirability Scale; FIML: Full information maximum likelihood; IM: Impression management; KSE-G: Social Desirability–Gamma Short Scale; MC-SDS: Marlowe-Crowne Social Desirability Scale; MG-CFA: Multi-group confirmatory factor analysis; MLR: Robust maximum likelihood; MMPI: Minnesota Multiphasic Personality Inventory; NQ–: Minimizing negative qualities; PQ+: Exaggerating positive qualities; SDE: Self-deceptive enhancement; SDR: Socially desirable responding; SDS-17: Social Desirability Scale-17; SES-17: Soziale-Erwünschtheits-Skala-17; UK: United Kingdom; Wsd: Wiggins Social Desirability Scale

Authors' contributions
All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]GESIS – Leibniz Institute for the Social Sciences, P.O. Box 12 21 55, 68072 Mannheim, Germany. [2]HSD University of Applied Sciences, Cologne, Germany.

References
Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). Boston, MA: Pearson.

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

Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion.* Chicago, IL: Rand McNally.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27,* 281–291. https://doi.org/10.1093/pubmed/fdi031

Brosseau-Liard, P. E., & Savalei, V. (2014). Adjusting incremental fit indices for nonnormality. *Multivariate Behavioral Research, 49*, 460–570. https://doi.org/10.1080/00273171.2014.933697

Brosseau-Liard, P. E., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate Behavioral Research, 47*, 904–930. https://doi.org/10.1080/00273171.2012.715252

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 464–504. https://doi.org/10.1080/10705510701301834

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

Colman, A. M. (2009). *A dictionary of psychology* (3rd ed.). Oxford: Oxford University Press. https://doi.org/10.1093/acref/9780199534067.001.0001

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

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

Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research, 47*, 615–639. https://doi.org/10.1177/147078530504700602

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. Westport, CT: Greenwood Press.

Haghighat, R. (2007). The development of the Brief Social Desirability Scale (BSDS). *Europe's Journal of Psychology, 3*. https://doi.org/10.5964/ejop.v3i4.417

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken, NJ: John Wiley & Sons.

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

Hathaway, S. R., & McKinley, J. C. (1951). *Minnesota Multiphasic Personality Inventory: Manual (Revised)*. San Antonio, TX: Psychological Corporation.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinions Quarterly, 67*, 79–125. https://doi.org/10.1086/346010

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling. A Multidisciplinary Journal, 6*, 1–55. https://doi.org/10.1080/10705519909540118

Kaminska, O., & Foulsham, T. (2013). *Understanding sources of social desirability bias in different modes: Evidence from eye-tracking* (ISER working paper series 2013-04). Retrieved from University of Essex Institute for Social and Economic Research website: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2013-04

Kemper, C. J. (2017). Content validity. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–4). Cham: Springer International Publishing.

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2014). Soziale Erwünschtheit-Gamma [KSE-G] [Social Desirability-Gamma Short Scale (KSE-G)]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen.* https://doi.org/10.6102/zis186

Kemper, C. J., & Menold, N. (2014). Nuisance or remedy? The utility of stylistic responding as an indicator of data fabrication in surveys. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 10*, 92–99. https://doi.org/10.1027/1614-2241/a000078

Kemper, C. J., Trapp, S., Kathmann, N., Samuel, D. B., & Ziegler, M. (2018). Short versus long scales in clinical assessment: Exploring the trade-off between resources saved and psychometric quality lost using two measures of obsessive-compulsive symptoms. *Assessment*. Advance online publication. https://doi.org/10.1177/1073191118810057

Li, A., & Bagger, J. (2006). Using the BDIR 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*, 131–141. https://doi.org/10.1111/j.1468-2389.2006.00339.x

Lück, H. E., & Timaeus, E. (1969). Skalen zur Messung manifester Angst (MAS) und sozialer Wünschbarkeit (SDS-E und SDS-CM) [Scales for the measurement of manifest anxiety (MAS) and social desirability (SDS-E and SDS-CM)]. *Diagnostica, 15*, 134–141.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates. https://doi.org/10.1111/j.2044-8317.1981.tb00621.x

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455. https://doi.org/10.1037/a0028085

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

Paulhus, D. L. (1986). Self-deception and impression management in tests responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire: Current issues in theory and measurement* (pp. 143–165). New York, NY: Springer. https://doi.org/10.1007/978-3-642-70751-3_8

Paulhus, D. L. (1988). *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding*. Unpublished manual, Department of Psychology, University of British Columbia, Vancouver.

Paulhus, D. L. (1991a). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). New York, NY: Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L. (1991b). Balanced Inventory of Desirable Responding (BIDR). In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 37–41). San Diego, CA: Academic Press.

Paulhus, D. L. (1998). *Paulhus Deception Scales (PDS): The Balanced Inventory of Desirable Responding–7: User's manual*. North Tonawanda, NY: Multi-Health Systems.

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

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic bias in self-perceptions: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality, 66*, 1024–1060. https://doi.org/10.1111/1467-6494.00041

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

Paulhus, D. L., Tanchuk, T., & Wehr, P. (1999, August). *Value-based faking on personality questionnaires: Agency and communion rule*. Paper session presented at the meeting of the American Psychological Association, Boston.

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

Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2018). Validation of the short and extra-short forms of the Big Five Inventory–2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000481

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184. https://doi.org/10.1177/01466216970212006

Roth, D. L., Snyder, C. R., & Pace, L. M. (1986). Dimensions of favorable self-presentation. *Journal of Personality and Social Psychology, 51*, 867–874. https://doi.org/10.1037/0022-3514.51.4.867

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research, 8*, 23–74.

Schweizer, K. (2010). Some guidelines concerning the modeling of traits and abilities in test construction. *European Journal of Psychological Assessment, 26*, 1–2. https://doi.org/10.1027/1015-5759/a000001

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Stöber, J. (1999). Die Soziale Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zu Reliabilität und Validität [The Social Desirability Scale-17 (SDS-17): Development and first results on reliability and validity]. *Diagnostica, 45*, 173–177. https://doi.org/10.1026//0012-1924.45.4.173

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70. https://doi.org/10.1177/109442810031002

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications to the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention:*

Methodological advances from alcohol and substance abuse research (pp. 281–324). Washington, DC: American Psychological Association.

Wiggins, J. S. (1959). Interrelationships among MMPI measures of dissimulation under standard and social desirability instruction. *Journal of Consulting Psychology, 23*, 419–427. https://doi.org/10.1037/h0047823

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

Wiggins, J. S. (2003). *Paradigms of personality assessment*. New York, NY: Guilford Press.