Social Desirability Bias in Attitudinal Measures

An Experimental Study Looking at Survey Modes, Respondent Traits and the Desirability of Survey Topics and Individual Questions

Henrik Kenneth Andersen

Chemnitz University of Technology

Faculty of Behavioural and Social Sciences

Institute of Sociology

Contact

Henrik.Andersen@soziologie.tu-chemnitz.de

+49 (0) 631 205 4971

1. Introduction

Social desirability (SD) bias describes respondents systematically presenting themselves in a more positive light than is accurate in self-reported surveys. It is a serious concern in survey research and can impact prevalence estimates of behaviour and attitudes as well as observed relationships between variables. While some researchers see SD bias as a result of a relatively stable personality trait present in some people, others emphasize more contextual elements including the sensitivity of the survey topics and items, the ambiguity (or lack thereof) of the social norms surrounding an attitude or behaviour as well as the reactions and consequences of their responses (Paulhus 2002; Stocké 2004b; Tourangeau & Yan 2007). Likely, it is at least to some extent a combination of both: some people feel a particularly strong need to present themselves positively, while considerations about the context and sensitivity of the situation moderates the extent to which this need manifests itself in survey responses (Stocké 2004b). While much research exists on the topic of SD and the mechanisms responsible for SD bias, empirical findings often turn up inconclusive, conflicting or even unexpected results.

This study takes another look at the interaction between survey mode and respondent traits, desirability and sensitivity. Importantly, it also closely examines an aspect that is often overlooked: the strength and direction of the desirability at not only topic- but also item-level. Contrary to self-reports involving more manifest sensitive behaviours, the use of expressions of opinions to measure latent attitudes can lead to a great deal of variability in desirability at the item-level.

Some work has been done to draw attention to this issue, most notable is perhaps that of Stocké & Hunkler (2007). They argue that the direction, strength and ambiguity (or lack thereof) of the social norms surrounding a survey item must be taken into account to properly interpret survey data.

Using multilevel modeling, both respondent-related and contextual aspects of the survey and the topics and questions included therein are analyzed to predict observed item scores. Typically, item scales intended to measure a latent construct are tested for their reliability and added or averaged to generate an aggregate attitude score for each respondent. The statistical analysis then often proceeds

with these aggregate measures. Multilevel modeling makes it possible to include information about the individual items that would typically be lost in a normal ordinary least squares (OLS) regression. The results of the empirical analysis suggest that SD bias can be the result of both respondent- and item-related characteristics. Furthermore, the anonymity of the situation seems to moderate some effects. Item-level desirability measures are helpful in predicting scores on their own and, in some cases, also in combination with survey mode. The results draw attention to the need to consider various sources of SD bias and cautions against making naive assumptions about the sensitivity and desirability of survey topics and items. Failure to do so may result in inaccurate conclusions.

The paper is structured as follows: in the first section, the motive for this paper and the general research question will be discussed. In the methods section, the measures used to assess attitudes towards the various out- or minority-groups will be presented. The main empirical analysis is a multilevel regression analysis with respondent and item (nested in topic) random effects and fixed effects for the experimental factor for survey mode and various respondent-, item- and topic-characteristics. Cross-level interactions between survey mode and the other fixed effects will be analyzed. Finally, in the concluding section, the main findings and their implications will be discussed.

2. Research Question

There is a large body of research examining the influence of survey modes on SD bias. Results of empirical analyses often show conflicting results. Generally, it has been assumed that anonymous surveys should reduce SD bias because they eliminate the motive for respondents to misrepresent themselves: they will not receive either approval or disapproval for their answers. However, results up until now have been mixed. While some studies have found evidence that anonymous surveys reduce SD bias (see for example Booth-Kewley et al. 2007; Kays et al. 2007; Bader et al. 2016; Dodou & de Winter 2014; Joinson 1999; Kreuter et al. 2008; Krysan et al. 1994), others (including some meta-studies) have shown evidence that this is not the case (Krysan 1998; Hancock & Flowers

2001; Börger 2013; Weisband & Kiesler 1996; Richman et al. 1999; Dwight & Feigelson 2000; Northover 2017, b; Rosenfeld et al. 1996; Holbrook et al. 2003).

This paper argues in support of the idea that anonymous surveys reduce SD bias but that it is necessary to control for other factors in order for the effect to become apparent. Specifically, it is argued that studies often fail to empirically test the desirability of the outcome variables and when they do, the level of specificity is not adequate. The study will look at the main question of how anonymity interacts with the desirability of ostensibly ‘sensitive’ topics as well as individual attitudinal indicators. As will be shown below, within what can be assumed to be ‘sensitive’ attitudinal topics (i.e. sexism, homophobia, islamophobia etc.), there is a great deal of variance in terms of desirability at the item-level.

Empirical studies looking at SD effects at both the topic- and item-level are relatively sparse, although Stocké (2004a) and Stocké & Hunkler (2007) had discussed the topic at some length. As Stocké wrote already in 2004, in order to properly assess SD bias one must look at the empirical desirability of the survey instruments. As to findings of no SD bias or SD bias in the opposite direction as hypothesized, he suggested: “This may be regarded as evidence that the respondents had desirability beliefs of different strength and direction in terms of the specific item content” (Stocké 2004a, p. 7). For these reasons, it is assumed that while anonymous surveys tend to lessen SD bias, the effect is likely contingent on the strength and direction of the perceived trait desirability of both the topics and items in question. In fact, this line of reasoning will inform the rest of the empirical analysis as well. Special attention will therefore be paid to the interaction of survey mode with respondent-, topic- and item-characteristics.

3. Method

To evaluate the influence of an anonymous interview setting on self-reports of prejudiced attitudes, an experiment was conducted. Respondents were randomly assigned to either a computer assisted personal-interview (CAPI) in which an interviewer posed questions and entered the respondents’ answers for them into a tablet computer, or a computer assisted self-interview (CASI) in which the

respondents were given the tablet computers to read and fill out the survey by themselves without revealing their answers to the interviewer.

The experimental survey was conducted in the German city of REMOVED FOR BLIND REVIEW (pop. roughly REMOVED FOR BLIND REVIEW) from April to May 2017. The survey was a non-probability sample of respondents over 16 with a complex-quota to reflect the city's makeup in terms of age, sex and district. In total, 372 interviews were conducted, of those 198 were in the CAPI mode, 174 in CASI. For the CAPI mode, 51.5% of the respondents were women (male: 48.5%) with a mean age for both sexes of 43.61 (std. dev.: 18.81), for CASI 48.9% were women (male: 51.1%) with mean age of 43.56 (std. dev.: 18.85).

3.1. Measures

Substantively, the survey focused on the overarching topic of attitudes towards various out- or rather minority-groups. The intercorrelation of various prejudices has been the topic of several studies under the banner of group-focused enmity (GFE) (Heitmeyer 2005; Wagner et al. 2008; Zick et al. 2008; Davidov et al. 2011). The study here concentrated on negative attitudes towards poor people, Muslims, women and homosexuals. Each of these latent constructs were measured each with 4-5 items. Items were measured using a 5-point Likert scale with 1: [the statement] “does not apply to me at all” to 5: “applies completely to me”. For most items, high scale values indicate prejudiced attitudes. In each construct, at least one of the items were formulated positively. For the following analyses, the positively worded items were first recoded to match the other items. For the outcome variable, higher scores always indicate more prejudiced attitudes. The items used come from both English and German-language scales (Atherton et al. 1993; Ashmore et al. 1995; Seise et al. 2002; Zick et al. 2008). The items displayed acceptable characteristics in the studies by the original authors and worked well in our study as well (see the descriptive statistics in Table 1 as well as the results of the confirmatory factor analysis (CFA) in Table 2). Other model variables include the respondents age, sex, authoritarianism (Schmidt & Berger 1995) and need for social approval (Crowne & Marlowe 1960).

3.2. Establishing Empirical Measures of Topic- and Item Trait Desirability

In accordance with recommendations from Stocké (2004a), it was decided not to measure the respondents' desirability beliefs directly but rather establish an aggregate trait desirability measure using a different sample. Doing otherwise may have also influenced the respondents' substantive answers on the outcome variables. Instead, a second sample based on the same quota-criteria was surveyed (n=66). Respondents were asked to evaluate how socially acceptable it is to express the attitudes and beliefs outlined in the main questionnaire items ("Item TD"). At the end of the survey, the respondents were asked to rate the generalized prejudiced attitudes (negative attitudes towards the poor, sexism, homophobia and islamophobia) as well ("Topic TD"). The scale ranged from -4: "societally very unacceptable" to +4: "societally very acceptable". The mean TD thus gives a measure of the direction and ambiguity (or lack thereof) of the norms surrounding a particular item or generalized attitude.

Table 1: Descriptive Statistics

Topic Alpha Mean (SD)	Item Mean (SD)	Text	Item TD Mean (SD)	Topic TD Mean (SD)
Poor 0.7 2.5 (0.8)	p01r	Society has a responsibility to help the disadvantaged. [+]	1.1 (1.9)	0.8 (1.7)
	p02	The government spends too much on welfare programs. [-]	0.1 (1.9)	
	p03	The disadvantaged are a burden on society. [-]	0.6 (1.8)	
	p04	Social assistance makes people lazy. [-]	1.3 (1.7)	
	p05	The disadvantaged are themselves responsible for their situation. [-]	0.6 (1.6)	
Muslims 0.8 2.4 (1.0)	m01r	Muslims should be able to live out their religious norms. [+]	-1.0 (2.1)	1.5 (1.9)
	m02	The many Muslims here make me feel like a foreigner in my own country. [-]	0.9 (2.0)	
	m03	The Muslims living here are a threat to our freedom and rights. [-]	0.5 (1.9)	
	m04r	Muslim women should be able to dress themselves according to their religious norms. [+]	-1.4 (1.9)	
	w01	The woman should be mainly responsible for the children and household. [-]	-0.3 (2.3)	
Women 0.8 1.8 (0.9)	w02	In marriage, the man should make the important decisions. [-]	-0.8 (2.3)	-0.6 (2.0)
	w03r	All types of jobs should be open to both men and women. [+]	2.1 (2.3)	
	w04r	Housework should be shared by men and women. [+]	1.2 (2.1)	
Homosexuals 0.9 1.9 (1.0)	h01	It makes me uncomfortable to see a gay/lesbian couple on the street. [-]	-0.0 (2.0)	0.2 (1.6)
	h02	A child's development suffers when it grows up with same-sex parents. [-]	0.7 (1.9)	
	h03	Homosexual men are less suited for leadership positions. [-]	-0.5 (2.0)	
	h04r	I could imagine being friends with a homosexual person. [+]	1.2 (1.9)	

As seen by the relatively large standard deviations of the trait desirability measures in Table 1, it is sometimes difficult to determine just how clear the social norm is surrounding a certain attitude; for some respondents, the trait may be considered desirable, for others undesirable. Thus, it is possible that mean trait desirability measures may not influence scores in a linear fashion. To take this into account, for each of the TD measures, a squared term was included into the model. This should reflect the fact that we should expect little effect from items or topics with mean trait desirabilities around the mid-scale range – here it is less certain whether the trait is desirable or

undesirable and should inform the respondents' desirability belief. As trait desirabilities increase in either the positive or negative direction, we should expect stronger effects.

In order to later properly interpret the effects of the trait desirability measures, it was necessary to also recode the ones associated with positively worded items. This is due to the fact that, in some cases, on average, respondents believed it was societally acceptable to hold prejudiced attitudes. Thus, we cannot naively assume that the socially desirable response is to downplay prejudice. In some cases, respondents wishing to gain approval should actually express more prejudiced attitudes. Without also recoding the mean TD scores appropriately, we would expect the TD as a predictor to exhibit a negative correlation with some items (and here some mental gymnastics are required: positive TD and positively worded (but recoded) item, i.e. this is the case for item p01r, refer to Table 1) and a positive correlation with others (positive TD and negatively worded (not recoded) item, i.e. p04).

Before conducting the rest of the empirical analyses, and in order to establish the construct validity of the measures, a CFA was done in Mplus 7.4 (Muthén & Muthén 1998–2018). In accordance with the findings of Heitmeyer (2005); Wagner et al. (2008); Zick et al. (2008); Davidov et al. (2011), the generalized nature of prejudiced attitudes was tested in a CFA in which the latent constructs for attitudes towards the poor, Muslims, women and homosexuals are seen as indicators for a the 2nd order latent factor for group-focused enmity (GFE). The CFA showed the measures are sound and the full results can be found Table 2.

3.3. Multilevel Models with Small Numbers of Clusters

The core of the empirical study are multilevel analyses in which scores on the attitude-items are explained using the experimental, respondent-, item- and topic-related predictors. The analysis was run using the lme4 package in R (Bates et al. 2015). The analysis is based on a 'less-is-better' assumption in which (after the necessary recoding) lower item scores (less prejudiced attitudes) indicate more SD-biased responses.²

Table 2: Confirmatory Factor Analysis

Latent Construct (1 st Order)	Latent Construct (2 nd Order)	unstd. est.	std. est.	s.e.
Poor	GFE	1.000	0.641	0.052
Women		1.237	0.747	0.044
Homosexuals		1.745	0.954	0.039
Muslims		0.850	0.534	0.051
Observed Variable	Latent Construct (1 st Order)	unstd. est.	std. est.	s.e.
p04	Poor	1.000	0.752	0.037
p02		0.979	0.736	0.043
p03		0.858	0.645	0.044
p05		0.737	0.554	0.044
w01	Women	1.000	0.881	0.024
w02		0.959	0.845	0.025
w03r		0.955	0.841	0.028
w04r		0.942	0.830	0.030
h01r	Homosexuals	1.000	0.767	0.033
h02		1.055	0.809	0.030
h03		1.111	0.852	0.030
h04r		0.850	0.628	0.043
m01r	Muslims	1.000	0.799	0.028
m02		1.060	0.847	0.029
m03		0.929	0.742	0.043
m04r		0.943	0.753	0.037

n=371, estimator: WLSMV, $\chi^2=271.701$, DF=99, p-value=0.000, CFI=0.963, TLI=0.955, RMSEA=0.069, C.I. RMSEA (90%)=0.059-0.078.

The data is structured as follows: each respondent (387) was asked to report their attitudes towards various out-groups (4) which were measured using opinion statements (17; 4 per topic, except for attitudes towards the poor, which was measured using 5). Items are nested in topics and items and topics are crossed with respondents. In order to account for the clustered nature of the data, a multilevel analysis was conducted. In contrast to typical OLS regressions, multilevel modeling allows us to consider multiple outcome variables simultaneously without eliminating crucial information about the individual items as is the case when running normal regression analyses with additive or averaged indices, for example. Based on the structure of the data, the most appropriate model seems to be one with random effects for respondents and items (nested in topics). To examine the extent to which respondent, item and topic characteristics interact with the survey mode, cross-level interactions will be specified. To establish the influence of the higher-level clusters on the outcome variables, the intraclass correlations (ICCs) were first examined in an empty model (model 0, not shown) which estimates the variance of the clusters (respondents, items nested in

topics) without any predictors at any level. The results suggest around 26% and 12% of the outcome variable's variance can be attributed to the respondent and the items, respectively. As such, this variance is not trivial (Hox 2010) and presents a methodological problem. It has been known for some time now that small sample sizes, especially concerning the number of clusters at level-two, can lead to inaccurate parameter estimates of fixed effects, variances and standard errors (McNeish & Stapleton 2016a,b; Stegmueller 2013; Hox et al. 2012). While some rules of thumb (i.e. 30 cluster/30 unit rule, see McNeish & Stapleton (2016a)) seem to have ingrained themselves, more recent work has drawn attention to various nuances. It is beyond the scope of this article to go into this in much more detail but, based on the given conditions, with 387 respondent and 17 item clusters, the extent of the bias should be close to or within nominal range (McNeish & Stapleton 2016a). Furthermore, several measures were taken to ensure the robustness of the model results including using a restricted maximum likelihood estimator (REML) to correct for variance component estimates and conducting a Kenward-Rogers (2009) adjustment to produce more conservative standard errors of the fixed effects.

All variables except for the dichotomous ones (mode, sex, positive wording) were centered before running the analysis. This was not done out of worries about multicollinearity but rather to assist with the interpretation of the coefficients (as some variables cannot empirically take on the value zero).

4. Empirical Analyses

Table 3 shows the full results of the multilevel analyses. The first model, with only the survey mode as a predictor, indicated that whether the survey was carried out as CAPI or CASI had no significant effect on item scores. This is in line with the majority of the literature cited above which also found no hard evidence for anonymous surveys significantly impacting SD bias.

The second model, which included the respondent-related characteristics shows several significant effects in the expected directions. Older respondents, males and those with authoritarian personalities all reported significantly more prejudiced attitudes. The respondent's need for social

approval based on the two-item modified Crowne and Marlowe scale had no significant effect. The respondent-related predictors lead to a decrease in the variance component of the random effects (see AIC, BIC, Log Likelihood and variance component estimates).

In model 3, the item-related characteristics were included. Here, the item's TD has a positive significant effect on item scores. The more socially acceptable the attitude is seen in society, the more respondents tended to report holding such an attitude. Again, because TD was measured using a bipolar scale and was recoded to fit the item wording, it means respondents tended to respond in a socially desirable way regardless of whether that meant expressing more or less prejudiced attitudes. We also see a noticeable reduction in the random effects variance component for items and topics but this will be discussed in detail below.

Finally, in model 4, the cross-level interactions were examined. The interaction between survey mode and sex (0.268, $p < 0.10$) shows that males reported more prejudiced attitudes in the anonymous CASI mode than CAPI while women, on the other hand, actually reported less prejudice in CASI. With several of the questions pertaining specifically to attitudes towards women and gender roles in general (including some of the items measuring homophobia), that men reported more positive attitudes in the CAPI mode seems understandable. A second marginally significant effect can be observed between mode and mean item trait desirability (-0.059, $p < 0.10$). The effect of trait desirability is slightly stronger for CAPI than for CASI. In non-anonymous surveys the respondents reacted more strongly to the (un)desirability of the items than when they could answer privately — it seems responses were more susceptible to desirability pressures in the open condition as opposed to the private one. As for the main-effects, the effects of age, authoritarianism and item TD remain significant and in the expected directions. Interestingly, the main-effect coefficient for survey mode actually becomes negative in model 4. The introduction of the interaction terms also substantially increases the standard error of some of the main effects. However, main-effects should be interpreted with caution as they are conditional on the scaling of the other variables. In interactions, main effects cannot be interpreted in terms of 'the effect of x when z or w is held constant' but rather 'the effect

of x contingent on the value of z (or w, etc.)'. Depending on the scaling of the other term in the interaction, it is possible to achieve very different main-effects (for more see Cohen et al. (2003); Jaccard & Turrissi (2003); Jaccard & Dodge (2004); Hayes (2013); Allison (2012); Mayerl & Urban (2018 expected)).

The respondent's need for social approval as measured by the modified short Crowne and Marlowe (CM) scale failed to show any significant effects on item scores. This is likely the result of the practical limitations of the survey as well as problems inherent to the scale (Paulhus 1986).

Table 3: Statistical Models

	Model 1	Model 2	Model 3	Model 4
(Intercept)	2.157*** (0.118)	2.099*** (0.121)	2.100*** (0.144)	2.164*** (0.147)
CASI	0.074 (0.072)	0.025 (0.065)	0.025 (0.065)	-0.113 (0.093)
Age		0.009*** (0.002)	0.009*** (0.002)	0.008** (0.003)
Male		0.176* (0.065)	0.176* (0.065)	0.050 (0.089)
Author.		0.253*** (0.036)	0.253*** (0.036)	0.248*** (0.047)
CM		-0.033 (0.045)	-0.033 (0.045)	-0.008 (0.062)
Positive			-0.131 (0.158)	-0.131 (0.158)
Item TD			0.312** (0.097)	0.340** (0.098)
Item TD ²			0.041 (0.081)	0.041 (0.081)
Topic TD			0.085 (0.124)	0.092 (0.125)
Topic TD ²			0.015 (0.157)	0.015 (0.157)
CASI*Age				0.002 (0.004)
CASI*Male				0.268 (0.131)
CASI*Author.				0.004 (0.072)
CASI*CM				-0.038 (0.090)
CASI*Item TD				-0.059 (0.034)
CASI*Topic TD				-0.016 (0.043)
AIC	18094.894	17890.093	17894.976	17923.676
BIC	18128.456	17950.413	17988.806	18057.720
Log Likelihood	-9042.447	-8936.047	-8933.488	-8941.838
Num. obs.	6078	6016	6016	6016
Num. groups:	370	362	362	362
Resp.				
Num. groups:	17	17	17	17
Topic:Item				
Var: Resp.	0.411	0.316	0.316	0.315
(Intercept)				
Var: Topic:Item	0.196	0.196	0.086	0.086
(Intercept)				
Var: Residual	1.001	1.007	1.007	1.007

***p < 0.001, **p < 0.01, *p < 0.05, ·p < 0.1

Estimator: restricted maximum likelihood (REML); Standard errors in parentheses; p-values calculated with Kenward-Roger adjustment; Analyses run using 'lme4' package in R (Bates et al. 2015)

The Model 3 results suggest a considerable portion of the estimated variance at the item and topic-level is explained by the higher-level predictors. The variance is reduced in Model 2 from 0.196 to 0.086 in Model 3. This reduction is mainly attributable to the single significant predictor Item TD. The more empirically socially acceptable (‘desirable’ even) a negative attitude is, the more prejudiced the responses were. Figure 2 shows the random effects estimates of items and topics of Model 1 compared to Model 4. The figure summarizes the substantive findings quite well and demonstrates the influence of social desirability on individual attitudinal items. While in Model 1 there was still considerable variability within and between the survey topics, the estimates begin to settle around the midpoint after controlling for the desirability of the items. This has several important implications. First, the concept of a generalized attitude of prejudice is supported. This was shown already in the CFA reported above. A substantial amount of the observed differences between items and topics can be traced back to social desirability bias. After controlling for TD, the differences are much less pronounced. The SEM model was seemingly able to largely correct the bias by identifying the underlying latent factor; the SD bias was captured by the error term. At the same time, the fact that survey mode effects are often not apparent in attitudinal scales now seems obvious. The variability of the item and topic random effects is largely not substantive but rather attributable to SD bias and, as such, attitude topics measured using opinion statements cannot be assumed to be uniformly desirable or sensitive.

5. Conclusion and Discussion

The results partially confirm what other researchers have found to varying degrees, namely that the survey mode alone does not seem to reliably affect responses and anonymous CASI surveys do not always reduce SD bias compared to CAPI. However, in conjunction with other factors, the context of the survey setting does seem to have an effect. Male respondents reacted strongly to the survey mode and gave markedly more desirable responses in the non-anonymous CAPI survey. As well, the interaction between survey mode and trait desirability shows respondents in the non-anonymous

CASI mode tended to modify their answers to conform with desirability concerns more so than in the CASI mode.

The findings point to the importance of looking at the many various sources of SD bias. As Stocké and his colleagues have argued, antecedents of SD bias, on their own, are often not sufficient. Without empirical information as to the strength, direction and ambiguity of the trait desirabilities, hypothesized SD effects are based on untested assumptions and are subject to confounding based on (possibly) heterogeneous desirability beliefs in the population. As was shown in this study, trait desirability was positively correlated with higher reported levels of prejudice. Incorporating the item TD into the model reduced the variance component of the random effects substantially. If it is seen as undesirable to hold negative attitudes, respondents will modify their responses to be more tolerant. If it is seen as desirable to hold negative attitudes, on the other hand, they will then modify their responses towards the more prejudiced.

Figure 2: Random Effects of Topic:Item

Graphics created using the 'sjPlot' package in R (Lüdtke 2017)

The cross-level interactions were shown to be only marginally significant. This was the result of a conscious choice to estimate more conservative standard errors to take the somewhat small number of clusters into account. As was discussed above, the model assumptions have likely not been

violated too gravely so this may have been an unnecessary precaution. Due also to the extensive discussion above and the fact that one could justify taking the one-tailed p-values for most of the estimates, I feel comfortable interpreting this result substantively, at least as a tentative direction for further research. In general, the results should be tested in more robust larger samples so as to increase the statistical power of the models.

The limitations of the study notwithstanding, it draws attention to the importance of looking at ostensibly sensitive attitudinal items individually. It does not seem helpful or wise to assume a certain topic is uniformly desirable or undesirable. Moreover, the very premise of sensitivity is largely contingent on the potential answers of the respondents. Based on the arguments laid out here, our hypothesis for the way in which SD bias will influence responses will be very atypical for a homophobe who believes it is societally accepted to express negative attitudes towards homosexuals. Likely, however, they will be more accurate than when blindly assuming that ‘everyone knows homophobia is bad’. Failure to take inter-item variability in terms of desirability into account may lead to a canceling-out of conflicting effects at the aggregate level. This may be in part responsible for the conflicting findings concerning how survey modes influence SD bias.

Notes

¹It should be noted that some authors distinguish between and use separate terms depending on whether the item desirability was measured according to the respondents' own beliefs about how desirable an attitude or behaviour is to *them* or concerning what they believe to be the prevailing opinion amongst their friends, in their neighbourhood or in society in general, see again Wolter (2012).

²This is a commonly used practice when looking at SD bias without validation data. In most experimental research, higher prevalence rates of sensitive behaviour in anonymous modes would be evidence enough to reject the null-hypothesis. However, as Stocké & Hunkler (2007) suggest, this may naively lead to incorrect conclusions without first testing the empirical desirability of the outcome variable.

References

- Allison, P. (2012). *When can you safely ignore multicollinearity?* Retrieved 2018-01-16, from <https://statisticalhorizons.com/multicollinearity>.
- Ashmore, R., Del Boca, F., & Bilder, S. (1995). Construction and validation of the gender attitude inventory, a structured inventory to assess multiple dimensions of gender attitudes. *Sex Roles*, 32(11-12), 753–785.
- Atherton, C., Gemmel, R., Haagenstad, S., Holt, D., Jensen, L., O'Hara, D. & Rehner, T. (1993). Measuring attitudes toward poverty: A new scale. *Social Work Research and Abstracts*, 29(4), 28–30.
- Bader, F., Bauer, J., Kroher, M., & Riordan, P. (2016). Privacy concerns in responses to sensitive questions. a survey experiment on the influence of numeric codes on unit nonresponse, item nonresponse, and misreporting. *methods, data, analyses*, 10(1), 47–72.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01.
- Blasberg, S., Rogers, K., & Paulhus, D. (2014). The bidimensional impression management index (bimi): measuring agentic and communal forms of impression management. *Journal of Personality Assessment*, 96, 523– 531.
- Booth-Kewley, S., Larson, G., & Miyoshi, D. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior*, 23(1), 463–477.
- Börger, T. (2013). Keeping up appearances: Motivations for socially desirable responding in contingent valuation interviews. *Ecological Economics*, 87, 155–165.
- Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd edition)*. Hillsdale: Erlbaum.
- Crowne, D., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349–354.

- Davidov, E., Thörner, S., Schmidt, P., Gosen, S. & Wolf, C. (2011). Level and change of group-focused enmity in germany: unconditional and conditional latent growth curve models with four panel waves. *Advances in Statistical Analysis*, 95(4), 481–500.
- Dodou, D., & de Winter, J. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, 36, 487–495.
- Dwight, S., & Feigelson, M. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement*, 60(3), 340–360.
- Hancock, D., & Flowers, C. (2001). Comparing social desirability responding on world wide web and paper-administered surveys. *Educational Technology Research and Development*, 49(1), 5–13.
- Hayes, A. (2013). *Introduction to mediation, moderation, and conditional process analysis. a regression-based approach*. New York: Guilford.
- Heitmeyer, W. (2005). *Gruppenbezogene Menschenfeindlichkeit. Die theoretische Konzeption und empirische Ergebnisse aus 2002, 2003, und 2004* (W. Heitmeyer, Ed.). Frankfurt: Suhrkamp.
- Holbrook, A., Green, M., & Krosnick, J. (2003). Telephone versus face-toface interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67(1), 79–125.
- Hox, J. (2010). *Multilevel analysis. techniques and applications. second edition*. New York: Routledge.
- Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a bayesian perspective. *Survey Research Methods*, 6(2), 87–93.
- Jaccard, J., & Dodge, T. (2004). *Analyzing contingent effects in regression models* (M. Hardy & A. Bryman, Eds.). Los Angeles: Sage.
- Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression (2nd edition)*. Newbury Park: Sage.

- Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods, Instruments and Computers*, 31(3), 433–438.
- Kays, K., Gathercoal, K., & Buhrow, W. (2012). Does survey format influence self-disclosure on sensitive question items? *Computers in Human Behavior*, 28, 251–256.
- Kenward, M., & Roger, J. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis*, 53, 2583–2595.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in cati, ivr, and web surveys. *Public Opinion Quarterly*, 72(5), 847–865.
- Krysan, M. (1998). Privacy and the expression of white racial attitudes. a comparison across three contexts. *Public Opinion Quarterly*, 62, 506–544.
- Krysan, M., Schuman, H., Scott, L.J., & Beatty, P. (1994). Response rates and response content in mail versus face-to-face surveys. *Public Opinion Quarterly*, 58, 381–399.
- Lüdtke, D. (2017). sjplot: Data visualization for statistics in social science [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=sjPlot>.
- Mayerl, J., & Urban, D. (2018 expected). *Mythen und methodische Probleme bei der Schätzung von Interaktionseffekten in linearen Regressionsanalysen*. Wiesbaden: VS Springer Verlag.
- McNeish, D., & Stapleton, L. (2016a). The effects of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28, 295–314.
- McNeish, D., & Stapleton, L. (2016b). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, 54(4), 495–518.
- Muthén, L., & Muthén, B. (1998–2018). *Mplus user's guide. sixth edition*. Los Angeles: Muthén and Muthén.
- Näher, A.-F., & Krumpal, I. (2012). Asking sensitive questions: the impact of forgiving wording and question context on social desirability bias. *Quality and Quantity*, 46, 1601–1616.
- Northover, S., Pedersen, W., Cohen, A., & Andrews, P. (2017). Artificial surveillance cues do not increase generosity: Two meta-analyses. *Computers in Human Behavior*, 38, 144–153.

- Paulhus, D. (1986). *Self-deception and impression management in test responses* (A. Angleitner & J. Wiggins, Eds.). Berlin: Springer.
- Paulhus, D. (2002). *Socially desirable responding: The evolution of a construct* (H. Braun, D. Jackson, & D. Wiley, Eds.). Hillsdale: Lawrence Erlbaum.
- Richman, W., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A metaanalytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires and interviews. *Journal of Applied Psychology*, 84(5), 754–775.
- Rosenfeld, P., Booth-Kewley, S., Edwards, J., & Thomas, M. (1996). Responses on computer surveys: Impression management, social desirability, and the big brother syndrome. *Computers in Human Behavior*, 12(2), 263–274.
- Schmidt, P., & Berger, M. (1995). *Stabilität und Wandel des Autoritarismus* (G. Lederer & P. Schmidt, Eds.). Opladen: Leske + Budrich.
- Seise, J., Banse, R., & Neyer, F. (2002). Individuelle Unterschiede im impliziten und expliziten Einstellungen zu Homosexualität. *Sexualforschung*, 15, 21–42.
- Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and bayesian approaches. *American Journal of Political Science*, 57(3), 748–761.
- Stocké, V. (2004a). Determinants and consequences of survey respondents' social desirability beliefs about racial attitudes. *Sonderforschungsbereich 504, Rationalitätskonzepte, Entscheidungsverhalten und ökonomische Modellierung*, 04–39, 1–31.
- Stocké, V. (2004b). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. Ein Vergleich der Prognosen der RationalChoice Theorie und des Modells der Frame-Selektion. *Zeitschrift für Soziologie*, 33(4), 303–320.
- Stocké, V. (2007a). Determinants and consequences of survey respondents' social desirability beliefs about racial attitudes. *Methodology*, 3(3), 125–138.
- Stocké, V. (2007b). The interdependence of determinants for the strength and direction of social desirability bias in racial attitude surveys. *Journal of Official Statistics*, 23(4), 493–514.

- Stocké, V., & Hunkler, C. (2007). Measures of desirability beliefs and their validity as indicators for socially desirable responding. *Field Methods*, 19(3), 313–336.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859–883.
- Wagner, U., Christ, O., & Pettigrew, T. (2008). Prejudice and group-related behavior in Germany. *Journal of Social Issues*, 64(2), 403–416.
- Weisband, S., & Kiesler, S. (1996). Self disclosure on computer forms: Meta-analysis and implications. *Proceedings of the SIGCHI Conference on Human Factors in Computer Systems*, 3–10.
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.
- Wolter, F. (2012). *Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik*. Wiesbaden: Springer VS.
- Zick, A., Wolf, C., Küpper, B., Davidov, E., Schmidt, P., & Heitmeyer, W. (2008). The syndrome of group-focused enmity: The interrelation of prejudices tested with multiple cross-sectional panel data. *Journal of Social Issues*, 64(2), 363–383.