Do Phantom Questions Measure Social Desirability?

Axel Franzen & Sebastian Mader
University of Bern

Abstract

Social desirability is a major problem in survey research. One way of handling the problem is to measure social desirability and to incorporate it into the statistical analysis. There are different techniques of measuring social desirability. We investigate and compare the performance of the well-known Crowne-Marlowe scale with the less common use of phantom questions. Up to now, there is only one study, which tests the comparative performance of both instruments (Randall & Fernandes 1991). In this paper we replicate the test and introduce a few innovations. In difference to the former study, we compare two short versions of the Crowne-Marlowe scale, the 10-items version as suggested by Clancy and Gove (1974) and a 10-items version suggested by Stocké (2014). First, we test both scales with respect to their internal consistency. Second, we investigate which of the two versions has the strongest impact on different sensitive behaviors (alcohol consumption, shoplifting, law compliance, and reported life satisfaction). Third, we construct 20 phantom questions, 10 with fictitious answering categories that can hardly be confused with existing things, and 10 where the fictitious categories resemble existing persons or sites. We then investigate whether the phantom questions pick up social desirability better than the Crowne-Marlowe scale. The study was conducted online with 365 student subjects. Our results indicate that the short version of the Crowne-Marlowe scale suggested by Clancy and Gove (1974) performs best. But none of our phantom questions or any combination of them is able to pick up social desirability. Instead over-claiming is associated with a lack of knowledge.

Keywords: social desirability, phantom questions, overclaiming, sensitive behavior, Crowne-Marlowe Scale

Social desirability is a major problem in survey research. Respondents usually have more or less the desire to report their true attitudes and behavior. However, when questions relate to sensitive topics they are also ashamed of reporting the true values and adapt their response towards what they believe is socially accepted or expected. This social desirability bias is well known and there are many examples of it in the literature (e.g. Tourangeau & Yan 2007; Wolter 2012). Very prominent examples stem from research about voting behavior or sexual behavior. For instance, the General Social Survey (GSS) asks men and women in the US for the number of sexual partners during their lifetime. Men report an average of 12.3 and women of 3.3 (Smith 1992). Similar results are obtained for Great Britain, France or New Zealand (Wiederman 1997). Assuming that both groups have roughly the same size and that sex involves usually one man and one woman the average must be the same. Hence, either men vastly exaggerate the true number or women reduce it or both. Also surveys about the participation in the last election or referendum usually generate much larger numbers than the known voting participation (Belli et al. 2001). There are many other examples that relate to tax evasion (Korndörfer et al. 2014) and other types of deviant behavior (e.g. Preisendörfer & Wolter 2014).

Basically, there are three ways of dealing with the social desirability bias. First, one possibility is obviously to not use surveys in sensitive research areas or at least to complement survey data with other observational or process generated data. A second strategy is to increase the anonymity of respondents. Besides using closed envelopes or question wording (which is actually not increasing anonymity but downplaying the sensitivity of the questions) anonymity can be increased by using self-administered interviews, or implementing special techniques like the randomized response technique (RRT) or related approaches like the crosswise model, or the item count model (ICT). The existing evidence suggests that self-administered interviews are less prone to socially desirable response behavior than personal interviews (Tourangeau & Yan 2007). Recent research on using RRT, the crosswise model or ICT suggests that they do not perform very well in surveys (e.g. Coutts & Jann 2011; Holbrook & Krosnick 2010; Höglinger et al. 2016; Wolter & Preisendörfer 2013). Often the sensitive behavior under investigation (e.g. plagiarism, or shoplifting) is lower when using these techniques as compared to direct questioning. Furthermore, a paper by Höglinger and Diekmann (2017) suggests that the "more is better" assumption does not always hold. In their study the number of participants who reported to have a very rare disease was higher using the crosswise model technique and hence further away from the true value than without

Direct correspondence to
Axel Franzen, University of Bern, Institute of Sociology, Fabrikstrasse 8, 3012 Bern, Switzerland
E-mail: franzen@soz.unibe.ch

this technique. A study by Höglinger and Jann (2018) compares different versions of RRT (forced-response and unrelated-question) and the crosswise model with respect to direct questioning and respondents' known behavior. They also report false positives using the crosswise model, and none of the RRT implementation outperforms direct questioning. One problem with indirect question techniques is that respondents do not understand the mechanism and react with high suspicion or increased random answering behavior.

A third strategy to deal with social desirability is to measure it. This was already suggested by Crowne and Marlowe in 1960. The original Crowne-Marlowe scale consists of 33 items that describe extreme behaviors or attitudes that hardly always apply to a respondent. An example is the item "I have never intensely disliked anyone". Respondents are then asked whether this statement describes their behavior or attitude as "true" or "false". The more "true" answers are given to socially desirable behaviors (as the example) or "false" answers to undesirable behaviors the higher is a respondent's score on the social desirability scale. The Crowne-Marlowe scale is the most applied measure of social desirability in survey research. Already Phillips and Clancy (1972) found that respondents scoring high on the Crowne-Marlowe (CM) scale also report higher overall life happiness (for similar results see Kozma and Stones 1987, Carstensen and Cone 1983) or report to have more friends as compared to respondents with low CM values. However, there is also some counterevidence. For example Johnson et al. (2012) find no association between the CM scale and cocaine use underreporting or with actual cocaine use as assessed by respondents' hair, saliva or urine samples. One problem with measures of social desirability like the CM scale is that it lacks "true" scores. Hence, it could be the case that the scale does not measure over- or underreporting but respondents' true behavior or attitude, or at least a mixture of both (Tourangeau & Yan 2007). One way to circumvent this problem is to use phantom questions. Such questions were already used by Phillips and Clancy (1972) in order to validate the social desirability scale of Crowne and Marlowe (1960, 1964). Phantom questions ask respondents whether they are familiar with certain people, books, movies or sites that do not exist. Hence, as opposed to the items used in the CM scale the true values of phantom questions are known and respondents are clearly overclaiming when responding to be familiar with non-existing people or sites.

So far there is only one study which compares the performance of the CM scale with the performance of phantom questions (Randall & Fernandes 1991). Randall and Fernandes (1991) use the full 33-items Crowne-Marlowe scale and five phantom questions that relate to consumer goods (movies, products, music albums, TV programs, and designer labels). The sensitive behavior under study referred to ten different acts of self-reported student misconduct (e.g. having plagiarized a term paper, turning in the same paper for two classes, cheating in exams). They find a negative correlation (*r* = - 0.24) between the Crowne-Marlowe scale and students' misconduct and no statistically significant correlation for the phantom questions. However, none of both measures were significant in the final multiple OLS regression analysis in which the authors included also a measure for the self-rated desirability of the ten sensitive behaviors in question. Consequently, the authors conclude "that further use of the M-C scale is not advisable" (ibid. 814). Similar conclusions apply to the phantom questions.

However, these conclusions are disputable. Randall and Fernandes (1991) measure trait desirability by asking respondents how desirable they believe each behavior under investigation is. These item-specific ratings are correlated with the CM scale (ibid. 811). From a theoretical perspective it is reasonable to assume that respondents' general measure of social desirability (CM scale) affects the desirability of specific behaviors (and not the other way round). For example, respondents' rating of the desirability of shoplifting could be influenced by the general tendency to answer in a socially desirable way. Under this assumption, trait desirability is a mediator variable and should not be included in one multiple regression model investigating the relation of the CM scale on sensitive behavior. Doing so wipes out (over-controlling) the correlation between the two (Morgan & Winship 2008, p. 65). Hence, the study might not be a reliable test of the performance of the CM scale.

Our study differs in a number of respects from the former study by Randall and Fernandes (1991): First, we refrain from including trait desirability for the reason already outlined above. Second, Randall and Fernandes (1991) use the full 33-items CM scale. However, this is a very long instrument and impractical for general population surveys. Therefore, we use two short versions of the CM scale, which are often found in the literature (Clancy 1971; see also Clancy & Gove 1974; Stocké 2014), and compare them with respect to their dimensionality, internal consistency, and their performance. Third, Randall and Fernandes (1991) use ten specific sensitive questions that only relate to typical student behavior like cheating in exams. In difference, our study includes questions from different areas such as respondents' level of norm compliance, alcohol consumption (Welte & Russell 1993; Embree & Whitehead 1993), shoplifting, and life satisfaction (Kozma & Stones 1987). Fourth, one reason why respondents might claim familiarity with non-existent objects or people in phantom questions might be the confusion with existing things. To study the impact of the confusion potential of phantom questions we designed one version having little confusion potential and one with a larger potential, and split the sample in such a way that every group received five phantom questions of each type.

The remainder of the article is structured as follows: Section two describes the two short versions of the CM scale and discusses their measurement characteristics. Section three presents the 20 phantom questions and contrasts the ones with and without the risk of confusion. Section four compares the criterion-related validity of the CM scales with the performance of the phantom questions. The final section concludes and discusses the results.

The Crowne-Marlowe Scale

To study the characteristics and performance of the CM-scale in comparison to phantom questions we conducted an online survey among the student population of the University of Bern. For this purpose, we randomly selected 2000 email addresses from the student email register and sent them an email including a link leading to the online survey in the beginning of March 2017. Overall, 463 students participated in the survey, which constitutes a response rate of 23.2%. The questionnaire contained about 70 questions, including 18 items of the Marlowe-Crowne scale, 10 phantom questions, and various questions on sensitive topics such as attitudes towards norm compliance, shoplifting, alcohol consumption, and life satisfaction. The median completion time of the survey was about 14 minutes. We excluded 70 participants from further analyses since their completion time was below 50% or above 200% of the median completion time. The rationale for doing so is that answering 70 questions in 7 minutes properly is probably not possible. Also, using 28 minutes for a 14 or 15-minute survey seems suspicious and might be due to respondents' attempt to search or google for true answers. We excluded an additional 28 respondents since they answered a test item instructing them not to provide any answer. The item reads "In this question we show you four answer categories. Please do not check any of the provided answer categories." Dropping cases with either a very short or a very long completion time and those with an invalid answer to the test item left us with 365 valid cases. However, the exclusion of these cases did not change any of the results substantially.

The original CM-scale consists of 33 items. Since this is a rather large instrument for a general survey most authors have used a reduced version of the CM scale. A prominent example is the 10-item short version suggested by Clancy 1971 (see also Clancy & Gove 1974; Phillips & Clancy 1972). Another short version was suggested by Stocké (2014). First, we discuss some measurement qualities of both scales separately. Second, we investigate whether the measurement qualities can be improved by some combination of both scales. The 10 items suggested by Clancy (1971) are shown in Table 1.

First, the distribution of CM1 is very close to normal and only slightly skewed to the left (skewness = -.13). Second, an exploratory principle component analysis (PCA) extracts four factors consisting of one or three items each. Third, Cronbach's alpha is .39 suggesting that the short version has low internal consistency. Both latter characteristics suggest that the items of Clancy's short version are rather heterogeneous. Next, we compare the short version suggested by Clancy (CM1) to a

Table 1 The short CM-Scale of Clancy 1971 (CM1)

		Polarity	CM1
I	(1) I never hesitate to go out of my way to help someone in trouble.	T	.73
	(2) On occasion I have had doubts about my ability to succeed in life.	F	.75
	(3) I sometimes try to get even, rather than forgive and forget.	F	.52
II	(4) No matter who I'm talking to, I'm always a good listener.	T	.63
	(5) At times I have really insisted on having things my own way.	F	.65
	(6) I have never been irked when people expressed ideas very different from my own.	T	.67
III	(7) If I could get into a movie without paying and be sure I was not seen, I would probably do it.	F	.69
	(8) There have been times when I felt like rebelling against people in authority even though I knew they were right.	F	.76
	(9) I never resent being asked to return a favour.	T	.24
IV	(10) Before voting I thoroughly investigate the qualifications of all the candidates.	T	.62

Note: N = 365, min = 0, max = 10, mean = 5.4, median = 5, modus = 5, sd = 1.92, Cronbach's alpha = 0.39. The numbers in the last column indicate factor loadings of a varimax rotated exploratory principle component factor analysis for polychoric correlations on components I, II, III, and IV, respectively. A component is identified if eigenvalue > 1. An equivalent analysis with simple Pearson's correlations yields substantially similar results. Each item has two answer categories, true (T) and false (F). Respondents receive an additional score for each item answered in the direction of social desirability, for example answering T to the first question or answering F to the second question.

different version suggested by Stocké (2014) (hereafter CM2). Also, Stocké (2014) picked 10 items from the original list (Table 2). Eight items differ from the CM1 version but two items appear in both short versions. These are the item number 5 of the Stocké version ("No matter who I'm talking to, I'm always a good listener.") and item number 7 ("Before voting I thoroughly investigate the qualifications of all the candidates."). Also, CM2 is almost normally distributed (skewness = -.20) and an exploratory factor analysis extracts also four components. Cronbach's alpha of CM2 is 0.53 and, hence, slightly better than the internal consistency of CM1 but still unsatisfactory.

Since both short versions have undesirable measurement qualities, e.g. multidimensionality and low consistency, we combined both scales to a 16-items ver-

Table 2 The short CM-Scale of Stocké 2014 (CM2)

		Polarity	CM2
I	(1) I sometimes feel resentful when I don't get my way.	F	.70
	(2) I'm always willing to admit it when I make a mistake.	T	.71
	(3) I am sometimes irritated by people who ask favors of me.	F	.46
	(4) I have never deliberately said something that hurt someone's feelings.	T	.64
II	(5) No matter who I'm talking to, I'm always a good listener.	T	.64
	(6) I am always courteous, even to people who are disagreeable.	T	.91
III	(7) Before voting I thoroughly investigate the qualifications of all the candidates.	T	.79
	(8) I keep getting myself on principles whose observance I expect from others.	T	.58
IV	(9) I can remember "playing sick" to get out of something.	F	.81
	(10) There have been occasions when I took advantage of someone.	F	.64

Note: N = 365, min = 0, max = 10, mean = 5.35, median = 5, modus = 6, sd = 2.09, Cronbach's α = 0.53. The numbers in the last column indicate factor loadings of a varimax rotated exploratory principle component factor analysis for polychoric correlations on components I, II, III, and IV, respectively. A component is identified if the eigenvalue > 1. An equivalent analysis with simple Pearson's correlations yields substantially similar results. Each item has two answer categories, true (T) and false (F). Respondents receive an additional score for each item answered in the direction of social desirability, for example answering F to the first question or answering T to the second question.

sion (CM3).[1] This 16-items version is depicted in Table 3. The CM3 version of the social desirability scale has a Cronbach's alpha of 0.62 and therefore, outperforms the CM1 and CM2 versions. However, like the other two short versions the scale is not one-dimensional but consists of five components as indicated by a principal component analysis (PCA).

[1] Two items were dropped since their inclusion resulted in lower Cronbach's alpha values.

Table 3 A composite scale of social desirability (CM3)

		Polarity	CM3
I	(1) There have been times when I felt like rebelling against people in authority even though I knew they were right.	F	.72
	(2) I sometimes try to get even, rather than forgive and forget.	F	.64
	(3) I have never deliberately said something that hurt someone's feelings.	T	.69
II	(4) If I could get into a movie without paying and be sure I was not seen, I would probably do it.	F	.59
	(5) Before voting I thoroughly investigate the qualifications of all the candidates.	T	.56
	(6) There have been occasions when I took advantage of someone.	F	.54
	(7) I keep getting myself on principles whose observance I expect from others.	T	.69
III	(8) At times I have really insisted on having things my own way.	F	.71
	(9) I sometimes feel resentful when I don't get my way.	F	.69
	(10) I am always willing to admit it when I make a mistake.	T	.37
	(11) I am sometimes irritated by people who ask favors of me.	F	.57
IV	(12) No matter who I'm talking to, I'm always a good listener.	T	.60
	(13) I have never been irked when people expressed ideas very different from my own.	T	.68
	(14) I am always courteous, even to people who are disagreeable.	T	.76
V	(15) I never hesitate to go out of my way to help someone in trouble.	T	.70
	(16) On occasion I have had doubts about my ability to succeed in life.	F	.65

Note: N = 365, min = 0, max = 16, mean = 8.50, median = 9, modus = 7, sd = 2.97, Cronbach's α = 0.62. The numbers in the last column indicate factor loadings of a varimax rotated exploratory principle component factor analysis for polychoric correlations on components I, II, III, IV, and V respectively. A component is identified if the eigenvalue > 1. An equivalent analysis with simple Pearson's correlations yields substantially similar results. Each item has two answer categories, true (T) and false (F). Respondents receive an additional score for each item answered in the direction of social desirability, for example answering F to the first question or answering T to the thrid question.

Phantom Questions

The CM-scale has the disadvantage that it lacks true values. Hence, it might not only pick up social desirability but also some true personality differences between respondents. This confusion might be one reason why the scale is multidimensional without any clear evidence why some items fall into one and others into another component. One alternative to measure social desirability or the need for social approval are phantom questions. Such questions ask respondents whether they are familiar with some objects, places or personalities that do not exist. The idea is that subjects with a strong need for social approval have a higher chance to claim that they are familiar with the person or object even if it does not exist since admitting not knowing something might create social disapproval. An example of a phantom question would be: "In the following we list four important international organizations. Which of these organizations do you know?" which is then followed by four answer categories "UNO", "OECD", "WIO", and "NATO". Obviously, "WIO" does not exist. But the answer has one problem. It is very close to "WHO" and thus, respondents might claim familiarity with WIO because they confuse it with WHO. Because of this risk of confusion, and because there is generally very little experience with phantom questions we generated 10 different phantom questions from various areas such as politics, geography, literature, architecture, science, movies, or generally concerning publicly known personalities. Additionally, we created two versions of every phantom question, one version of which we thought that the risk of confusion is low and one in which it is higher. Generally, the risk of confusion is higher if the fictitious issue sounds similar to an existing issue in contrast to an issue that is distinct from an existing site or person. All twenty questions are listed in Table 4. Because it would be cumbersome for respondents to answer twenty such knowledge questions in one survey we randomly split the questionnaire in two versions. Version one contained the first five phantom questions without the risk of confusion and the last five with the risk of confusion. The other version was the other way around and contained the first five phantom questions with the risk of confusion and the last five without risk of confusion. This way we had two groups of respondents who answered each ten phantom questions. This design enables us to study the effect of low or high risk of confusion on the answering behavior of phantom questions.

After a short introduction, the online questionnaire started with the 10-items CM scale of Clancy and Gove (1974), followed by five phantom questions, continued with 10-items of the CM scale of Stocké (2014), and was again followed by the remaining five phantom questions. Questions on more or less sensitive opinions and behaviors followed in the middle. The questionnaire concluded with sociodemographic information. Each block of phantom questions was split into two parts showing three phantom questions on the first screen and two on the following screen.

Table 4 displays the proportion of respondents who answered “yes” to the four categories of the 10 phantom questions. We are interested here in the proportions of “yes” answers to fake categories. Table 4 shows that the proportion of yes-answers to fake items without risk of confusion varies between 0% and 9% and is therefore relatively low. None of the respondents said that they are familiar with an Oscar – winning movie called “sense of delight” and 9% thought that Peter Dickens was an American President. The proportions of yes answers are considerably higher when the fake answer is formulated in such a way that the risk of confusing it with existing places or people is higher. Proportions vary between 5% and 49% with the risk of confusion. 5% of respondents said that they are familiar with a Nobel Prize Winner called Jassir Peres, and 49% claimed that they are familiar with an architectural style called “futurism”. The proportions of yes-answers are consistently higher when we purposely tried to increase the risk of confusion. Hence, this intended manipulation worked quite well. However, surprisingly phantom questions do not correlate very high among each other. This is true for the first five phantom questions without risk of confusion and the last five one with risk of confusion in group one (highest $r = .50$) as well as for those phantom questions in group 2 (highest $r = .28$). Practically none of our respondents consistently claimed familiarity with all fictitious items.

This already points into the direction that phantom questions are very context specific but do not pick up consistently a personality trait such as the need for social approval. Furthermore, there are also no obvious sequence effects. Phantom questions were presented in the order displayed in Table 4. Only 1% answered that they are familiar with EBO (first item), 2% with the author Jean-François Le Gouguec, 6% with Sevenstone Cave, 7% with Modular Style, and 3% with the Fun Loving Animals. Hence, there is no indication of learning effects, such that respondents improved their performance with the number of phantom questions. Similar observations apply to the sequence of the other phantom questions.

Comparing the Criterion-Related Validity of the CM-Scale with the Performance of Phantom Questions

The questionnaire contains a number of questions on sensitive topics, such as whether respondents ever took something from a store without paying for it (shoplifting), how many glasses of alcohol they consume during a week, whether they believe that laws should always be adhered, and on their general life satisfaction.

Table 4 Description of the Phantom Questions by risk of confusion (ROC)

		Without ROC		With ROC	
I	International Organizations: In the following, we list four important international organizations. Which of these organizations do you know?	UNO	(98)	UNO	(99)
		OECD	(70)	OECD	(68)
		EBO	(1)	WIO	(8)
		NATO	(99)	NATO	(100)
II	Authors of Universal Literature: In the following, we list four authors of Universal Literature. Which of these authors do you know?	Johann Wolfgang von Goethe	(99)	Johann Wolfgang von Goethe	(99)
		William Shakespeare	(100)	William Shakespeare	(100)
		Mark Twain	(91)	Mark Twain	(88)
		Jean-François Le Gouguec	(2)	Niki de Saint Phalle	(24)
III	UNESCO World Heritage Sites: Now we list four UNESCO World Heritage Sites. Which of these sites do you know?	Venice	(99)	Venice	(100)
		Sevenstone Cave	(6)	The Mexican Wall	(19)
		Pyramids of Gizeh	(90)	Pyramids of Gizeh	(80)
		Taj Mahal	(92)	Taj Mahal	(92)
IV	Architectural Styles: Now we list four architectural styles. Which of these epochs do you know?	Bauhaus	(52)	Bauhaus	(46)
		Jugendstil	(85)	Jugendstil	(86)
		Gothic Style	(98)	Gothic Style	(97)
		Modular Style	(7)	Futurism	(49)
V	Environmental Protection Organizations: In the following, we list four important international environmental protection organizations. Which of these organizations do you know?	Fun Loving Animals	(3)	World Climate Protection Trust	(9)
		United Nations Environmental Programme	(20)	United Nations Environmental Programme	(20)
		World Wildlife Fund	(64)	World Wildlife Fund	(67)
		Greenpeace	(100)	Greenpeace	(100)

Table 4 continued

		Without ROC		With ROC	
VI	Famous Musicians: Which of the following four musicians do you know?	Kurt Cobain	(91)	Kurt Cobain	(95)
		Paul McCartney	(95)	Paul McCartney	(94)
		Sandy Lawn	(2)	Bob Cohen	(21)
		Amy Winehouse	(100)	Amy Winehouse	(99)
VII	US Presidents: In the following, we list four former US Presidents. Which of these politicians do you know?	Barack Obama	(100)	Barack Obama	(100)
		Peter Dickens	(9)	George Adam	(10)
		John F. Kennedy	(100)	John F. Kennedy	(100)
		Abraham Lincoln	(98)	Abraham Lincoln	(98)
VIII	Charitable Celebrities: In the following, we list four famous personalities that take a stand for charity. Which of these persons do you know?	George Clooney	(99)	George Clooney	(100)
		Angelina Jolie	(99)	Angelina Jolie	(100)
		Gabriele Goldau	(2)	Jenifer Cruz	(20)
		Bill Gates	(99)	Bill Gates	(99)
IX	Oscar-winning Movies: In the following, we list four Oscar-winning movies. Do you know these films?	Sense of Delight	(0)	A Beautiful Girl	(8)
		Gladiator	(91)	Gladiator	(89)
		Titanic	(99)	Titanic	(98)
		Forrest Gump	(97)	Forrest Gump	(97)
X	Nobel Peace Prize Winners: Which of the four following Nobel Peace Prize Winners do you know?	Dalai Lama	(100)	Dalai Lama	(100)
		Michail Gorbatschow	(73)	Michail Gorbatschow	(73)
		Martin Luther King	(97)	Martin Luther King	(100)
		Aleksander Islic	(0)	Jassir Peres	(5)
Dummy for ≥ 1 'yes' answer (I)-(X) (n=365)			(12)		(56)

Note: Arabic numbers in parentheses indicate the percentage of 'yes' answers. Fake items are underlined. For questions (I)-(V) n = 174 for the fake items without ROC and n=191 with ROC. For questions (VI)-(X) n = 191 for the fake items without ROC and n = 174 with ROC.

All these questions were taken in the exact same formulation as they usually appear in large general population surveys. Agreeing to law compliance and life satisfaction are socially desirable matters. Respondents who are identified of having a high need for social approval should therefore more strongly over-report those behaviors as compared to individuals who care less about social approval. Results of multiple OLS regression analyses are displayed in Table 5. Every line of the table represents the results of an independent multiple regression model in which we control for all available socio- demographic variables (age, sex, subject of study, nationality, main language, household size, designated study degree). As can be seen, all CM scales are positively related to agreement with law compliance and life satisfaction as expected. Hence, the CM scale does pick up over-reporting. The effects of the CM2 and CM3 scales are a little weaker than the effects of the CM1 scale, and are statistically insignificant with respect to life satisfaction and norm compliance. Alcohol consumption and shoplifting should be underreported by respondents with a high need of social approval and this is what can be observed from the results of Table 5. Here, all three CM versions perform equally well. Respondents with a high need for social approval report to drink less alcohol and report less often that they have shoplifted before (logit model).

Next, we investigate how the phantom questions perform. For this purpose, we constructed two different scales. Respondents are coded as being sensitive towards social desirability if they have claimed familiarity with at least one or more fake sites, objects, organizations, or persons when the risk of confusion was low (without ROC), and when the risk of confusion was high (with ROC). The results of both versions are displayed in lines 4 and 5 of Table 5. As can be seen from the results neither version is statistically significantly associated with any of the four dependent variables. These results are robust if we ran 20 models including each time a different phantom question or if the index is composed of both versions of phantom questions (with and without the risk of confusion), or if the index is constructed continuously by summing up the number of wrongly answered phantom questions.

Hence, phantom questions do obviously not measure social desirability. This raises the question of what phantom questions measure instead. One obvious answer is that they simply measure knowledge. We therefore conducted a second study trying to find evidence for this explanation. The second study was conducted in May 2017 at the University of Bern with N = 318 respondents. The original purpose of the second study was to investigate the relation of IQ test scores (see Liepmann et al. 2012) with emotional intelligence and empathy. However, the online questionnaire which respondents had to answer in the laboratory contained also some of the same phantom questions used in Study 1 (questions I, II, III, VI and IX without risk of confusion). The relevant results of Study 2 are displayed in Table 6. The dependent variable is the dichotomous characteristic of whether respondents answered

Table 5 Regressions of various traits on CM and overclaiming

Model		(1)	(2)	(3)	(4)
		OLS	OLS	OLS	Logit
Dependent Variable (z-stand.)		Law compl.	Happiness	Alcohol	Shoplifting
CM1 (z-stand.)		0.11*	0.22***	-0.11*	-0.50***
		(0.06)	(0.05)	(0.05)	(0.12)
	adjusted R^2	0.03	0.05	0.08	
	pseudo R^2				0.08
CM2 (z-stand.)		0.08	0.12	-0.12*	-0.38**
		(0.05)	(0.06)	(0.05)	(0.12)
	adjusted R^2	0.02	0.02	0.08	
	pseudo R^2				0.07
CM3 (z-stand.)		0.10	0.18**	-0.10*	-0.45***
		(0.05)	(0.06)	(0.05)	(0.12)
	adjusted R^2	0.02	0.04	0.08	
	pseudo R^2				0.07
Overclaiming (without ROC)		-0.00	-0.21	-0.12	0.31
		(0.16)	(0.17)	(0.11)	(0.34)
	adjusted R^2	0.01	0.01	0.07	
	pseudo R^2				0.04
Overclaiming (with ROC)		-0.07	0.07	0.00	-0.40
		(0.11)	(0.11)	(0.11)	(0.24)
	adjusted R^2	0.01	0.00	0.07	
	pseudo R^2				0.05
n		348	348	348	347

Note: Displayed are the standardized regression coefficients. * = p<0.05, ** = p<0.01, *** = p<0.001. All standard errors (in parentheses) are robust with respect to heteroscedasticity. All models control for sex, age, German mother tongue, Swiss nationality, designated study degree, household size, and study subject. Table A1 summarizes the descriptive statistics of all variables in the models. Note that all results remain robust even if respondents with a very low or very high completion time remain in the sample.

Table 6 Logistic Regression of Overclaiming on IQ

Model Dependent Variable (z-stand.)	(1) Overclaiming
IQ (z-stand.)	-0.65*** (0.19)
Female	-0.07 (0.39)
Age	-0.03 (0.11)
Mother Tongue: German	0.40 (0.68)
Swiss	-0.03 (0.81)
Household Size	-0.04 (0.17)
Designated degree: Master	0.05 (0.66)
University of Bern	-0.03 (0.44)
Constant	-1.59 (2.91)
n	297
pseudo R^2	0.05
Loglikelihood	-98.53

Note: Displayed are logit coefficients. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. All standard errors (in parentheses) are robust with respect to heteroscedasticity.

one or more phantom questions wrongly.[2] The logistic regression results of Table 6 indicate that subjects with high IQ test scores claimed statistically significantly less often of being familiar with non-existing things, objects, or people. None of the other control variables (sex, age, subject of study and so on) were found to be associated with overclaiming of phantom questions. This is also true for the same

[2] The dependent variable has values ranging from 0 (no claim of familiarity with any fictitious item) to 5 (claiming familiarity with each fictitious item). Of all respondents 37 (12%) stated to be familiar with at least one fictitious item, and eleven respondents with more than one. Giving this skewed distribution, we dichotomized the dependent variable and used a logistic regression. However, using a negative binomial model gives the same results.

analysis of the data from the first study. Taken together, our results suggest that phantom questions measure knowledge but not the need for social approval.

Summary and Discussion

A comparison of the performance of three different short versions of the CM scale with respect to self-reports on law conformity, shoplifting, alcohol consumption, and life satisfaction suggests that the CM scale picks up social desirability. As expected, higher values on the CM scale are positively associated with opinions on law compliance and life satisfaction. The standardized coefficients show that the effect sizes are small. Furthermore, all three versions detect also underreporting of shoplifting and alcohol consumption as expected. Moreover, our study shows that it basically does not matter whether we use the short version suggested by Clancy (1971), or a combined version of the Stocké and Clancy scale with 16 items. The combined version has a higher Cronbach's alpha value but the associations with sensitive behavior are almost the same as with the CM1 scale. Hence, our study confirms the finding of other studies suggesting that the CM scale works.

However, we did not find a single association with one or any combination of phantom questions with sensitive behavior (shoplifting, alcohol consumption, norm compliance, life satisfaction). Also, phantom questions have small correlations among each other and no correlation with any short version of the CM scale (see Table A2). These results suggest that phantom questions measure knowledge but not the need for social approval. Of course, our study results are obtained from a student sample which raises questions on the generalizability. However, limitations of generalizability mainly apply to descriptive results but less to associational findings. Theoretically, it is possible that phantom questions pick up social desirability in a general population sample but not in a student sample. However, practically this is very unlikely.

In contrast to phantom questions, we find that all three versions of the CM scale are associated with the sensitive behaviors studied, and that the CM1 version outperforms the other two versions slightly. This finding might suggest, that the CM scale measures social desirability. However, the finding is also compatible with the interpretation that the CM scale as well as the sensitive behavior(s) are both caused by true but unobserved personality differences. In that case, the correlation between the CM scale and the desirable behavior in question would be spurious. This omitted variable bias can only be avoided in validation studies in which the true behavior of respondents is known. Such studies are rare. One recent study by Preisendörfer and Wolter (2014) does not find a statistically significant relation between the CM scale and truthful answering whether respondents have been convicted of a crime. However, also Preisendörfer and Wolter (2014) included trait desirability in their analysis together with the CM scale and, therefore, might have introduced an over-control bias into their study. Hence, further research on the validity of the CM scale and improvements on measuring social desirability are still in need.

References

Belli, R. F., Traugott, M. W., & Beckmann, M. N. (2001). What leads to voting overreports? Contrasts of overreporters to validated voters and admitted nonvoters in the American National Election Studies. *Journal of Official Statistics*, 17(4), 479-498.

Carstensen, L. L., & Cone, J. D. (1983). Social desirability and the measurement of psychological well-being in elderly persons. *Journal of Gerontology*, 38(6), 713-715.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research*, 40(1), 169-193.

Clancy, K. (1971). Systematic bias in field studies of mental illness. Ph.D. dissertation, New York University.

Clancy, K., & Gove, W. (1974). Sex Differences in Mental Illness: An Analysis of Response Bias in Self-Reports. *American Journal of Sociology*, 80(1), 205-216.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology,* 24(4), 349-354.

Crowne, D., & Marlowe, D. (1964). The Approval Motive. Studies in Evaluative Dependence. New York: John Wiley & Sons.

Embree, B. G., & Whitehead, P. C. (1993). Validity and reliability of self-reported drinking behavior: Dealing with the problem of response bias. *Journal of Studies on Alcohol*, 54(3), 334-344.

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Survey Research Methods,* 10(3), 171-187.

Höglinger, M., & Diekmann, A. (2017). Uncovering a Blind Spot in Sensitive Question Research: False Positives Undermine the Crosswise-Model RRT. *Political Analysis,* 25(1), 131-137.

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

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly,* 74(2), 328-343.

Johnson, T. P., Fendrich, M., & Mackesy-Amiti, M. E. (2012). An evaluation of the validity of the Crowne-Marlowe need for approval scale. *Quality and Quantity,* 46(6), 1883-1896.

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax evasion: Improving self-reports using the crosswise model. *Journal of Economic Psychology*, 45, 18-32.

Kozma, A., & Stones, M. J. (1987). Social desirability in measures of subjective well-being: A systematic evaluation. *Journal of Gerontology,* 42(1), 56-59.

Liepmann, D., Beauducel, A., Brocke, B., & Nettelnstroth, W. (2012). Intelligenz-Struktur-Test. Screening. Göttingen: Hogrefe.

Morgan, S. L., & Winship, C. (2008). Counterfactuals and Causal Inference: Methods and Principles for Social Research. New York: Cambridge University Press.

Phillips, D. J., & Clancy, K. J. (1972). Some effects of „social desirability" in survey studies. *American Journal of Sociology,* 77(5), 921-940.

Preisendörfer, P., & Wolter, F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly,* 78(1), 126-146.

Randall, D. M., & Fernandes, M. F. (1991). The Social Desirability Response Bias in Ethics Research. *Journal of Business Ethics,* 10(11), 805-817.

Smith, T. (1992). Discrepancies Between Men and Women in Reporting Number of Sexual Partners: A Summary from four Countries. *Social Biology,* 39(3-4), 203-211.

Stocké, V. (2014). Deutsche Kurzskala zur Erfassung des Bedürfnisses nach sozialer Anerkennung. Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS, doi:10.6102/zis159.

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin,* 133(5), 859-883.

Welte, J. W., & Russell, M. (1993). Influence of socially desirable responding in a study of stress and substance abuse. *Alcoholism – Clinical and Experimental Research,* 17(4), 758-761.

Wiederman, M. W. (1997). The Truth Must Be in Here Somewhere: Examining the Gender Discrepancy in Self-Reported Lifetime Number of Sex Partners. *The Journal of Sex Research*, 34(4), 375-386.

Wolter, F. (2012). Heikle Fragen in Interviews. Eine Validierung der Randomized Response-Technik. Wiesbaden: Springer VS.

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research*, 42(3), 321-353.

Appendix

Table A1 Descriptive statistics of all dependent and control variables of Table 5

	Variable	mean	sd	min.	max.	n	Question wording
Dependent Variables	Norm compliance	3.64	1.07	1	5	361	"Laws should always be complied to, no matter how agreeable they are." Five answering categories from 1 'disagree strongly' to 5 'agree strongly'.
	Happiness	7.42	1.63	0	10	361	"All in all how satisfied are you with your life?" 11-point Likert scale ranging from 0 'very unsatisfied' to 10 'very satisfied'.
	Alcohol Consumption	2.98	3.58	0	25	369	"How many glasses of wine, beer, or other alcoholic beverages do you drink in a usual week?" Number of weekly glasses of wine, beer, or other alcoholic beverages.
	Shoplifting	.37		0	1	366	"Have you ever in your life taken something deliberately from a store without paying for it?" Answer categories 1 'yes' and 0 'no'.

Table A1 continued

	Variable	mean	sd	min.	max.	n	Question wording
Control Variables	Sex: Female	.60		0	1	360	Dummy, 1 if female
	Age	24.73	4.79	19	59	360	in years
	German Mother Tongue	.92		0	1	360	Dummy, 1 if yes
	Swiss Nationality	.91		0	1	360	Dummy, 1 if yes
	Household Size	3.19	1.18	1	5	360	Number of people living in the household
	Bachelor	.62		0	1	358	Dummy, 1 if designated study degree is Bachelor (reference category)
	Master	.33		0	1	358	Dummy, 1 if designated study degree is Master.
	Ph.D.	.05		0	1	358	Dummy, 1 if designated study degree is Ph.D.
	Study Subject:						
	Economics and Social Sciences	.34		0	1	369	Dummy, 1 if yes (reference category)
	Law	.13		0	1	369	Dummy, 1 if yes
	Natural Sciences	.15		0	1	369	Dummy, 1 if yes
	Medicine	.14		0	1	369	Dummy, 1 if yes
	Humanities	.24		0	1	369	Dummy, 1 if yes

Table A2 Correlation matrix of the different CM scales and overclaiming scales (phantom questions)

	CM1	CM2	CM3	Overclaiming without ROC	Overclaiming with ROC
CM1					
CM2	0.59***				
CM3	0.86***	0.87***			
Overclaiming without ROC	-0.00	-0.02	-0.00		
Overclaiming with ROC	0.00	0.05	0.00	0.20***	

Note: Displayed are Pearson's correlation coefficients. *** = p < 0.001.