Erschienen in: Personality and Individual Differences ; 107 (2017). - S. 164-173
https://dx.doi.org/10.1016/j.paid.2016.11.023

An item level evaluation of the Marlowe-Crowne Social Desirability Scale using item response theory on Icelandic Internet panel data and cognitive interviews

Vaka Vésteinsdóttir [a,*], Ulf-Dietrich Reips [b], Adam Joinson [c], Fanney Thorsdottir [d]

[a] *University of Iceland, Department of Psychology, Aragata 14, 101 Reykjavik, Iceland*
[b] *University of Konstanz, Department of Psychology, Fach 31, 78457 Konstanz, Germany*
[c] *University of Bath, School of Management, Bath, UK*
[d] *University of Iceland, Department of Psychology, Sturlugata 3, 101 Reykjavík, Iceland*

A B S T R A C T

The Marlowe-Crowne Social Desirability Scale (MCSDS) is commonly used to validate other self-report measures within social and health research. Concerns over the scale's length (33 items) have repeatedly been raised. Nevertheless, prior efforts to develop psychometrically sound short forms of the MCSDS have not led to consistent findings. The purpose of this study was to develop a short form of the MCSDS, in accordance with guidelines for best practices in short form and scale development. Information on item properties, obtained with item response theory (IRT) and cognitive interviews (CogI), were used to eliminate items with poor properties and select items for a short form to be administered via the Internet. The IRT analyses were based on responses from 536 Internet panel members and the CogI sample consisted of 40 interviewees. Ten items were dropped due to poor psychometric properties and out of the 23 remaining items a ten item short form was developed.

Keywords:
Marlowe-Crowne Social Desirability Scale
Short form
Psychometric properties
Cognitive interviews
Item response theory
Mixed methods

1. Introduction

The Marlowe-Crowne Social Desirability Scale (MCSDS) (Crowne & Marlowe, 1960) is a 33 item commonly used instrument for measuring social desirability response style (SDRS). SDRS is a respondent's tendency to present him- or herself favorably and can have confounding effects on self-reported data (see e.g. Kaufmann & Reips, 2008; Podsakoff et al., 2003; Podsakoff et al., 2012). It is therefore important, when using self-reports, to be able to obtain estimates of SDRS. However, adding a 33 item scale to a questionnaire places extra burden on respondents, potentially preventing the use of the MCSDS in research when cost and/or respondent fatigue are of major concern. The aim of this study is therefore to extend the Vésteinsdóttir et al. (2015) study of the psychometric properties of the MCSDS, by evaluating single items from the MCSDS for the purpose of eliminating items with the weakest psychometric properties and suggesting a psychometrically sound short form of the scale.

1.1. Social desirability response style

Evidence consistent with SDRS comes from studies on self-reports of behaviors such as illicit substance use, alcohol use, smoking, abortion, energy consumption, income, criminal behavior, voting behavior, exercise and seat belt use (Tourangeau & Yan, 2007). SDRS is a serious problem in assessment because it can inflate the scores on desirable items and deflate the scores on undesirable items, resulting in biased estimates and possible distortion of relationships between variables. SDRS can therefore have a significant confounding effect on empirical findings and lead to misleading conclusions (see Cote & Buckley, 1988). Concerns about the effect of response biases on the validity of research findings have been shown to influence how reviewers perceive the quality of results and subsequent decisions about publication of work (Pace, 2010).

If the target variables are not related to a measure of SDRS, it can be concluded that they are free of SDRS and that their relationships are not distorted by the bias. If, however, SDRS is identified in the target variables, researchers must control for the effects of the bias in subsequent analysis (for an overview see for example Podsakoff et al., 2003). There are a number of statistical techniques for controlling for the effects of SDRS but the recommended one is to use a direct measure of SDRS in a latent variable model (Podsakoff et al., 2012).

* Corresponding author.
E-mail addresses: vakav@hi.is (V. Vésteinsdóttir), reips@uni-konstanz.de (U.-D. Reips), A.Joinson@bath.ac.uk (A. Joinson), fanneyt@hi.is (F. Thorsdottir).

1.2. The Marlowe-Crowne Social Desirability Scale

The MCSDS has been extensively used to validate target measures and control for the effect of SDRS. Between 1960 (when the MCSDS was first published) and 2002, >1000 articles and dissertations mentioned the use of the MCSDS when the PsychINFO, ERIC, Sociological Abstracts and Social Sciences Abstracts databases were searched (Beretvas et al., 2002) and according to Google Scholar in May 2016, >7000 works had cited the original article on MCSDS. The number of studies using the original MCSDS, or some of its short forms, continues to grow to this day (see van Schie et al. (2016) for a recent example of the use of the MCSDS, and Black and Reynolds (2016) for recent use of a MCSDS short form, in Internet administrated scale validation).

Some have, however, questioned the use of SDRS measures such as the MCSDS (see e.g. Tracey, 2016) due to a controversy over the scale's content and dimensionality (Barger, 2002; Fischer & Fick, 1993; Helmes & Holden, 2003; Leite and Beretvas, 2005; Loo & Thorpe, 2000; McCrae & Costa, 1983; Paulhus, 1984; Ventimiglia & MacDonald, 2012). Other researchers have opted for the BIDR (Paulhus, 1991) as a measure of SDRS, partly because of the often presumed two dimensional nature of SDRS (see e.g. Davis et al., 2012) and because the BIDR was developed with newer and more sophisticated techniques (Lambert et al., 2016). New research has, however shown that the MCSDS is unidimensional in an Internet administration (Vésteinsdóttir et al., 2015) and outperforms the BIDR in detecting faking (Lambert et al., 2016), suggesting both adequate psychometric properties and usefulness of the MCSDS.

A major limitation of the MCSDS is the length of the scale. The MCSDS consists of 33 true/false items, which describe behaviors that are "culturally sanctioned and approved of but which are improbable of occurrence" (Crowne & Marlowe, 1960, p. 350). Adding 33 items to an assessment procedure places extra burden on respondents and adds to the cost of administration. This is particularly true of instruments such as the MCSDS which are used for validation of other assessment tools and thus presented with at least one other measure. Increased length of a questionnaire can reduce potential participants' willingness to respond and increase the likelihood of exhausting respondents' patience (fatigue effect), which can result in non-completion of questionnaires and reduced response quality (Galesic & Bosnjak, 2009; Reips, 2010; Schuman & Presser, 1996). It would therefore be desirable to have a shorter measure of SDRS in order to have more space for questions on the assessment topic and reduce response burden and possible cost.

1.3. Short form development

For these reasons, researchers have attempted to develop short forms of the MCSDS scale, selecting items based on results from exploratory factor analysis (Ballard, 1992; Reynolds, 1982; Strahan & Gerbasi, 1972). Unfortunately, however, these attempts have not led to consistent findings (Vésteinsdóttir et al., 2015). Previous efforts have three main limitations: First, the short forms have been created using statistics which rely heavily on sample specific statistics. This is probably the main reason why previous attempts have not agreed on which items should be omitted from the short form. Secondly, the emphasis in previous studies was on selecting items for short forms to maximize internal consistency. However, focusing only on internal consistency in short form development may create a short form that is too narrow and potentially low in validity (Loevinger, 1954). Finally, short form developments have exclusively relied on convenience student samples (Vésteinsdóttir et al., 2015). Clearly, a student sample does not represent the population for which the scale is intended. The consequences of such sample non-representativeness can severely harm the short form development efforts (DeVellis, 2012). Furthermore, guidelines for best practices in scale and short form development recommend the use of multiple indicators of quality (Clark & Watson, 1995; DeVellis, 2012; Stanton et al., 2002).

One approach to overcome the limitations in previous short form development, listed above, is to use item response theory (IRT) to obtain information on item properties, instead of the previously employed techniques. The benefit of using IRT in short form development is that IRT models provide information on item properties in relation to respondents' estimated trait level (how much of the characteristic being measured, the respondent is presumed to possess). Taking respondents' trait level into account means that IRT estimates are not as highly dependent on the characteristics of the sample as methods that are purely based on item responses (see Embretson & Reise, 2000 for a more in depth explanation). The most commonly used short forms of the MCSDS have been developed using component factor analysis (Vésteinsdóttir et al., 2015), which is based on correlations between item responses. This method has the drawback of favoring redundant items (the more similar the items, the higher the correlation will be between them). With IRT the items are placed on a continuous scale, which represents the characteristic being measured. The items can thus be chosen to either measure as many points on the continuum as needed (e.g. when making a short form of a scale) or to have high precision at a certain point of the scale (see e.g. DeVellis, 2012; Embretson & Reise, 2000).

1.4. The present study

As the discussion above has highlighted, there is a need for a short and psychometrically sound version the MCSDS. In this study, items with the weakest psychometric properties will be identified and eliminated and a short form of the MCSDS (i.e. MCSD-SF) will be developed based on best practices in short form development. A combination of item response theory (IRT) and cognitive interviews (CogI) will be used to evaluate each item. In addition, previous factor analyses of the MCSDS will be included in the analysis to identify items that have repeatedly obtained the lowest factor loadings.

In view of the increasing number of studies that collect data online (e.g. Reips, 2012), the short form is intended for the Internet survey mode and thus the IRT analysis are based on Internet survey data with CogI, conducted face to face, for the purpose of evaluating the clarity of the items.

2. Method

2.1. Online survey

2.1.1. Participants and procedure

The IRT analysis was done on a sample collected by the Social Science Research Institute (SSRI) in July 2013 (see Vésteinsdóttir et al., 2015 for description). The survey, containing all 33 items of the MCSDS in Icelandic, was presented on three pages, each containing 11 items, to be consistent with other surveys sent out by the SSRI where lengthy question grids are generally avoided. An e-mail invitation was sent out to 1200 potential participants, drawn from the SSRI Internet panel. Duration of data collection was two weeks, with three reminders being sent out within the first 12 days after the original invitation was sent. Out of the 639 participants who responded to at least one item on the MCSDS, 536 participants (44.7% of the original sample) completed all items on the scale. Evaluation of psychometric properties of data obtained with the MCSDS was based on completed scales. The final sample consisted of 272 women (50.7%) and 264 men (49.3%), aged between 20 and 81 years (mean age being 49 years), with educational levels varying from elementary school education to a post-graduate university degree.

2.1.2. Instrument

The *Marlowe-Crowne Social Desirability Scale* (Crowne & Marlowe, 1960) contains 33 TRUE/FALSE items, 18 keyed in the true direction (attribution of desirable behavior) and 15 keyed in the reverse direction

(denial of undesirable behavior). Responses in the keyed direction are coded as one and responses in the opposite direction as zero. The maximum score on the MCSDS is therefore 33 and the minimum zero, with higher scores indicating more social desirability in responses. An Icelandic translation of the MCSDS (Vésteinsdóttir et al., 2015) was used in this study.

2.1.3. Data analysis

Little's MCAR test (Little, 1988) indicated that omitted responses were missing completely at random ($\chi^2 = 1497.77$, $df = 1478$, $p = 0.354$) and therefore a listwise deletion was used. Item response theory (IRT) was used to evaluate the MCSDS items. The IRT model is based on the assumptions that the scale in use is unidimensional and that items are locally independent. A scale is unidimensional if all items belong to the scale measure of one and the same construct. Local independence means that the items are independent given a respondent's score on the underlying latent trait (for further discussion see Hambleton et al., 1991). It should be noted, that although SDRS is being referred to here as a trait, this does not mean that SDRS is assumed to be stable across all situations, or much like Hambleton et al. (1991) noted in reference to the dominant factor as ability: "…ability is not necessarily inherent or unchangeable" (p. 10).

IRT was used to evaluate both item discrimination (a) and item sensitivity (b) – usually referred to as the item difficulty or threshold parameter. The fit of a unidimensional two parameter logistic (2PL) model was tested with Tay and Drasgow's (2012) adjusted χ^2/df ratio, using the Excel macro MODFIT (Stark, 2001). Tay and Drasgow (2012) proposed the adjusted χ^2/df with the χ^2 adjusted (for say n = 3000) to overcome problems related to the χ^2 sensitivity to sample size. For doubles and triples of items, ratios >3 are indications of model misspecification and thus a violation of the unidimensionality assumption. The mean adjusted χ^2/df will be presented with a cutoff value of 3. Model fit was further examined with fit plots, produced by MODFIT, showing the correspondents between the empirical and expected response functions.

A 2PL model was set up in R using the *ltm* package (Rizopoulos, 2006) and tested against a Rasch model for further justification of model choice. The difference between these models lies in the parameter estimation for each item. In the simplest of these models, the Rasch model, a is set equal for all items, estimating only the b parameter for all items, which is the point of inflection on the θ scale (Baker, 2001; Edelen & Reeve, 2007). The b parameter is typically said to denote the difficulty of the item, with low estimates indicating that an item is easily endorsed (or denied if negatively worded). This stems from the history of IRT model development, which occurred mainly within the fields of ability testing and education (see Embretson & Reise (2000) for a brief history of IRT) where the latent trait usually refers to some ability, which would enable the respondent to overcome the difficulty of getting an item "correct" (the keyed response being the correct response). However, in the current context θ refers to the tendency to respond in a socially desirable manner, i.e. the tendency to describe oneself in a certain way. It would seem unfitting to speak of the "difficulty" of giving a certain self evaluation and thus we will refer to the b parameter as an estimate of item sensitivity – with low estimates of b indicating that individuals with low scores on the underlying latent trait (θ) have a relatively high probability of responding in the keyed direction, meaning that the item is sensitive to SDRS. The reverse holds for high estimates on b. Estimates of b are (in most cases) standardized with an average of 0 and a standard deviation of 1 and will thus typically range between −2 and 2 (Reise & Henson, 2003).

In the 2PL model the item discrimination parameters (or slopes) are free to vary across items, allowing the estimation of a, which is the slope to a constant of the item characteristic curve (ICC, a function of the probability of a keyed response for a given level of θ) at the point of inflection. The a parameter indicates the item's ability to discriminate among individuals high and low on θ (and can range from positive to negative infinity, though negative discrimination will not be discussed here). The a parameter estimates can be approximated as follows: 0 no discrimination, 0.01–0.34 very low, 0.35–0.64 low, 0.65–1.34 moderate, 1.35–1.69 high, >1.70 very high and + infinity perfect discrimination (Baker, 2001), with typical values ranging between 0.05 and 1.5 (Reise & Henson, 2003). The suitability of the 2PL model was tested by comparing the 2PL model to the Rasch model by evaluating the change in fit using a Chi-square difference test based on the log-likelihoods of the nested models (LR test), with degrees of freedom equal to the difference in the number of parameters estimated for each model. The comparative fit indices Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are also given. Both indices are parsimony-adjusted with lower values indicating a better fit. However, the AIC and BIC do not always agree, as the BIC penalizes more for the number of parameters and thus may favor simple models more strongly than the AIC (Kang et al., 2005).

Confirmatory factor analysis (CFA) was used to confirm the one factor structure of the MCSDS short form, previously obtained for the full scale (Vésteinsdóttir et al., 2015). A one factor model was estimated with DWLS (diagonally weighted least squares) in the *lavaan* package in R (Rosseel, 2012). Factor loadings and random errors were estimated freely but the factor variance was set to one to enable identification of the model. Error variances were not allowed to covary. The fit indices used to evaluate the fit of the model estimated were: χ^2 (Chi-square), CFI (comparative fit index), TLI (Tucker-Lewis-Index, a.k.a. nonnormed fit index; NNFI) and RMSEA (root mean square error of approximation). CFI and TLI indicate whether the model fits the data better than a model that does not assume any relation between measured variables, while adjusting for sample size. Both of these indices range from zero to one. The cutoff criterion for an adequate fit was set at 0.96 for CFI and 0.95 for TLI. The RMSEA index indicates how well the model fits the data. Values of RMSEA range from zero to one, with a lower value signifying a better fit. The cutoff criterion for an adequate fit for RMSEA was set at 0.05. Fit indices and their cutoff criteria were chosen based on how they have performed when DWLS was used in medium-large samples (Yu, 2002). In addition, we checked the fit indices χ^2/df and RSMR (standardized root mean square residual). A χ^2/df below 2 indicates a good model fit (Bollen, 1989). For RSMR, which is a measure of the standardized difference between the observed and predicted correlations, values below 0.08 indicate acceptable fit (Hu & Bentler, 1999).

Raykov's (1997) composite reliability estimate (ρx) for the MCSDS short form and full scale are also calculated based on factor loadings obtained with CFA such as described above. The criterion for good reliability was set at ρx = 0.80 or higher (DeVellis, 2012), although it should be noted that this criterion is based on Cronbach's α which gives the lower bound of the reliability coefficient (see Raykov, 1997).

2.2. Cognitive interviews

2.2.1. Participants

Forty participants were recruited for the interviews, 20 men and 20 women. Participants were selected based on their age, with the aim of obtaining responses from all age groups, within the age range of Internet panel members (18 and older). The median age of participants was 45 years, ranging from 18 to 73 years.

2.2.2. Procedure

A protocol for probing interviews was prepared to maintain congruence between interviewers and ensure the compatibility of interviews. A series of questions about each statement of the Icelandic version of the MCSDS was created. The MCSDS statements were presented verbally in the same order as they appear in the original scale. Participants were asked to give a true/false response to the MCSDS statement presented, immediately followed by questions about the MCSDS statement/response process, before being presented with the next MCSDS statement. The interviews were structured in the sense that all probing

questions for each MCSDS statement had to be asked, but the interviewers could repeat and/or rephrase the questions, if participants' responses indicated any confusion about the task. An example of an MCSDS statement with probing questions is presented below. The first two and the last question presented in the example were the same for all the MCSDS items.

MCSDS 7. I am always careful about my manner of dress.

a. TRUE or FALSE?
b. What does it mean to say that the statement is TRUE/FALSE (depending on the respondents answer)?
c. How do you interpret *always* in this context?
d. What does it mean to *be careful* in this context?
e. How do you interpret *manner of dress* in this context?
f. Is there anything that you find confusing about this statement?

The interviews were conducted by one of two trained interviewers, in most cases in the interviewee's home. The interview process was explained to the participants prior to the interview. Participants were informed that the interview would be recorded and ensured that no one apart from the interviewer would have access to the original recording, which would be deleted immediately after it had been transcribed. Participants were also told that there were no right or wrong answers to the statements, their task would only be to answer the questions truthfully. Most of the interviews were completed in one session without a break, but a few participants requested a break during the interview. The interviews took approximately 1 h each.

2.2.3. Data analysis

All interviews were recorded and transcribed. A bottom up technique was used to identify and categorize problems related to each statement. This was done by going through the transcripts, making a note of each indication of a problem and grouping together problems of the same type to form categories. Three categories were found to be most descriptive of problems related to item clarity. The categories and their descriptions are as follow:

Understanding: The most common problem, identified with the cognitive interviews, was different understanding of words and terms (identified in 23 items). This can be a serious problem if the difference in understanding is such that the content of the item is changed and respondents are essentially answering different questions (Fowler, 1995). When measuring SDRS the main question is whether respondents' understanding of a statement alters the desirability of the behavior described in that statement. If so, the same response may not indicate the same level of SDRS.

Frequency: The second most common problem reported in the CogI was the use of frequency words (identified in 22 items). Specifically, when words and phrases that describe the frequency of behavior, feelings or thoughts appear in items their interpretation may vary by the context in which they are presented and the respondent's own experience. Even the words *always* and *never* (found in positive items), which have a clear meaning, are sometimes taken to mean something like *almost always* and *almost never*. In addition, interviewees often found it difficult to explain their interpretation of words and phrases that refer to a frequency that is in between never and always (which can only be found in negative items), and explanations of the same word/phrase varied from one interviewee to another. This was especially true of the word *sometimes* (in Icelandic: "stundum").

Presumptions: The third group of items that seemed problematic was those that described behaviors in certain situations, where the respondents' familiarity with that situation is presumed (identified in 4 items). However, if some respondents have no experience of the situation described (e.g. have never voted, do not own a car, etc.) the basis for their responses is not the same as for those familiar with the given situation. They could for example deny having done the behavior in question because they have never been in that specific situation – and not because it is something they wouldn't do, or something that would not be representative of their behavior in general. If however a presumption holds for the majority of survey respondent, this source of non-clarity is not likely to affect parameter estimates much. The size of the problem therefore depends on the proportion of a sample that the presumption does not hold for.

3. Results

3.1. Fit of the 2PL model

As can be seen in Table 1 the mean adjusted χ^2 for the doubles and triples did not exceed the cutoff value of 3, indicating a good fit of 2PL model.

An examination of fit plots for the 33 items also indicated good fit of all the items as the empirical item response function corresponded well with the expected item response function.

3.2. Comparison of nested models

The LR test showed that the 2PL model produced a significantly better fit than the Rasch model. The AIC and BIC indices were not in agreement as the AIC indicated a better fit for the 2PL model and the BIC indicated a better fit for the Rasch model (see Table 2). However as the BIC has a tendency to favor simple models and both the LR test and AIC favored the 2PL model, the 2PL model was chosen.

3.3. Item elimination: CogI and IRT analysis

Item parameter estimations for the 2PL model are presented in Table 3. Item 1 has the lowest item discrimination parameter with an estimate that falls within the cut criterion for a very low *a* value (0.01–0.34). Items: 2, 7, 8, 17, 18, 24, 25 and 29 all have low estimates of *a* (the cut criterion being 0.35–0.64 for low values), suggesting removal of these items. The discrimination parameter is a good indicator of the severity of the problem, but not for the reason of the problem. The reverse is true of the CogI, as they do not obtain measures of magnitude (only whether or not an item falls in a non-clarity category – not how often or how much), they do however, provide possible reasons as to why the discrimination parameter of an item is low. The following information on the main problems identified in the above mentioned items was obtained with CogI:

Item 1. Before voting I thoroughly investigate the qualifications of all the candidates. How the interviewees interpreted the words *investigate*, *qualifications* and *candidates* affected how they responded to the item. Different examples were given of the act of investigating qualifications by interviewees who gave a *TRUE* response, naming things like following what the candidates say publicly (on TV or in the newspapers), and by those who gave a *FALSE* response, describing a more in depth look at candidates' background (education, experience, employment record etc.). The interviewees who answered in the keyed direction also took the word *candidates* to mean parties or the leaders of the parties, not each individual candidate, making the task much easier. It was also noted by the interviewers that the interviewees who gave a *FALSE* response seemed to have no problem admitting that they did not perform the behavior in question.

Item 2. I never hesitate to go out of my way to help someone in trouble. When responding to this item, interviewees either thought of *someone* they know (friends or family members) or took *someone* to refer to a stranger. None of the respondents who answered *FALSE* to this statement thought of people they know. Another, and perhaps more serious, problem was the interpretation of the word *trouble*. The interviewees gave examples of *trouble* that varied in type (financial, physical and/or psychological problems) and severity (e.g. needing a lift or having suffered a car accident). Different combinations of the relation to the person in question, the type of trouble they are in and the severity of it

Table 1
Frequency table of adjusted (n = 3000) χ^2/df.

		<1	1 < 2	2 < 3	3 < 4	4 < 5	5 < 7	>7	Mean	SD
Adjusted χ^2 ratio	Singlets	33	0	0	0	0	0	0	0	0
	Doublets	445	19	16	8	7	10	23	0,978	3305
	Triplets	3952	358	305	217	156	187	281	1295	2,75

are numerous and lead to very dissimilar items, creating multiple possible reasons for either a *TRUE* or *FALSE* response.

Item 7. I am always careful about my manner of dress. Some interviewees took the word *always* to mean "always when around others" or "always except when at home". Others, however, took it to mean "without exception" – none of whom gave a *TRUE* response to the item. The interpretation of *manner of dress* could also vary from being decently dressed (not wearing worn-out and/or unclean clothes) to being fashionably dressed. There is quite a difference between being fashionably dressed all the time and being decently dressed most of the time, the latter being something that most people do, regardless of their level of social desirability.

Item 8. My table manners at home are as good as when I eat out in a restaurant. The word *restaurant* could refer to anything from a fast food restaurant or café to a fine dining restaurant, making the presumption invalid that people behave better in restaurants.

Item 17. I always try to practice what I preach. This is a commonly used phrase in English but because it is a phrase, its meaning goes beyond the content of the sentence and is therefore difficult to translate to a language that does not have a corresponding phrase. In the Icelandic version, the item only captures whether people try to behave consistently and in accordance with their life values, regardless of whether they preach these values to others, which is a much easier task.

Item 18. I don't find it particularly difficult to get along with loud-mouthed, obnoxious people. Interviewees found the item confusing and difficult to answer. The negation followed by a double-barreled ending of the statement made it difficult for them to respond to the item and many interviewees said that they could get along with *loud-mouthed* people but not *obnoxious* people and were therefore not sure how to answer the question.

Item 24. I would never think of letting someone else be punished for my wrongdoings. The severity of the *punishments* interviewees thought of varied from being told off to getting a jail sentence or even a death sentence, with those interviewees that gave a *FALSE* response thinking of milder punishments than those who gave a *TRUE* response. This is however a hypothetical question (as can be seen by the mentioning of death sentences, which are illegal and considered inhumane in Iceland and in almost all of Europe and thus a highly unlikely scenario for respondents), so it may be that unrealistic scenarios produce unrealistic responses. It should also be noted that one interviewee refused to answer the question saying that there was no way of knowing what one would do. Furthermore, two interviewees based their initial responses on what they would do, not on what they would *think of* doing –confusing the two scenarios.

Item 25. I never resent being asked to return a favor. Many of the interviewees who answered in the keyed direction changed the meaning of *never* to almost never, usually not, most likely not etc. making it easier to give a *TRUE* response to the item.

Item 29. I have almost never felt the urge to tell someone off. Only one of the interviewees answered in the keyed direction and therefore not much can be said about this item in terms of the non-clarity categories and keying. The one interviewee was however the only one to interpret *almost never* as seldom, whereas other interviewees took it to mean: never, once in a lifetime or very rarely. Overall, the interviewees interpreted the term *tell someone off* in different ways and some were unable to give an explanation of what the term meant.

The items with the highest discrimination estimates are items 6 and 28. High values of *a* indicate that the item makes a clear distinction between individuals with θ scores around the point of inflation on the ICC and are therefore very informative for given levels of θ, whereas items with very low discrimination provide little information because changes in θ produce little change in the probability of a keyed response. However a problem with using item discrimination to choose items is that most of the items with low estimates of *a* are positively worded and most of the items with high estimates are negatively worded (see Fig. 1). Selecting items solely on this criterion would therefore produce a short form with a disproportionate number of negative items. On a ten item short form, for example, only two of the items would be positively worded.

Table 2
Comparison of nested models.

	AIC	BIC	G^2	*df*	*p*-Value
Fit Rasch	19,877.62	20,019.0			
Fit 2PL	19,772.74	20,055.5	170.87	33	<0.001

Table 3
Item parameter estimation for the 2PL model, item non-clarity indicators identified with the cognitive interviews and low factor loadings obtained in previous studies.

Item	IRT		Cogl item non-clarity indicators	Factor loadings <0.30
	a	*b*		
1	0.33 (0.10)	1.76 (0.60)	UP	bc
2	0.57 (0.14)	−3.26 (0.77)	U	
3*	0.76 (0.12)	0.26 (0.13)	F	b
4	0.89 (0.13)	0.95 (0.16)		
5*	0.70 (0.13)	1.48 (0.27)	UF	b
6*	1.69 (0.22)	1.05 (0.11)	UF	
7	0.48 (0.11)	0.46 (0.21)	UF	abc
8	0.58 (0.11)	−0.08 (0.16)	UP	
9*	1.12 (0.18)	−1.84 (0.24)	UP	a
10*	0.87 (0.13)	0.31 (0.12)	UF	b
11*	0.98 (0.14)	0.83 (0.14)	UF	
12*	0.70 (0.12)	−1.23 (0.23)	UF	
13	1.03 (0.14)	0.71 (0.13)	F	
14*	1.08 (0.14)	−0.10 (0.10)	U	b
15*	1.33 (0.16)	0.06 (0.09)	UF	
16	1.05 (0.14)	−0.66 (0.12)	F	
17	0.57 (0.18)	−4.53 (1.35)	UF	c
18	0.41 (0.11)	1.99 (0.55)	U	abc
19*	1.31 (0.17)	−0.96 (0.12)	F	
20	0.92 (0.17)	−2.41 (0.38)		
21	0.83 (0.14)	−1.48 (0.23)	UF	
22*	0.65 (0.12)	1.18 (0.24)	U	a
23*	1.16 (0.15)	0.41 (0.10)	F	
24	0.57 (0.17)	−4.11 (1.12)	UF	c
25	0.61 (0.17)	−3.98 (1.04)	F	b
26	0.75 (0.12)	−0.26 (0.13)	UF	
27	0.74 (0.12)	−0.18 (0.13)	UP	ab
28*	1.48 (0.18)	0.52 (0.09)	UF	
29	0.47 (0.12)	2.72 (0.69)	UF	abc
30*	1.08 (0.14)	−0.02 (0.10)	F	
31	1.10 (0.15)	0.42 (0.11)	U	
32*	0.87 (0.13)	−1.01 (0.17)	UF	b
33	1.00 (0.14)	−0.16 (0.11)	F	

Note: *a* = discrimination parameters (standard errors in brackets), *b* = sensitivity parameters (standard errors in brackets). Items marked with * are keyed in the reverse direction. a = factor loadings below 0.30 in Reynolds (1982), b = factor loadings below 0.30 in Ventimiglia and MacDonald (2012), c = factor loadings below 0.30 in Vésteinsdóttir et al. (2015). Item clarity indicators: U = understanding, F = frequency, P = presumption.

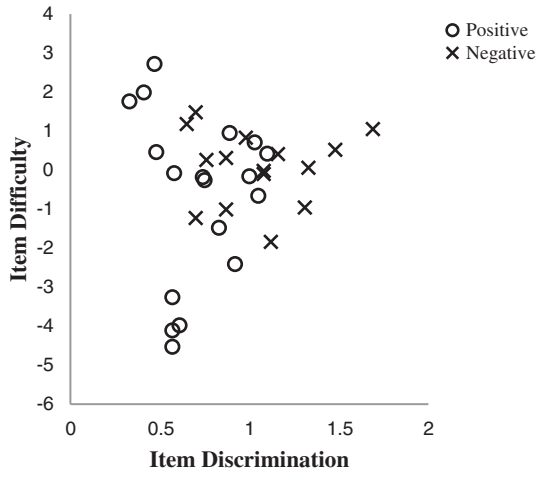


Fig. 1. Item discrimination estimates plotted against item sensitivity estimates for positively and negatively worded items.

The original scale contains more positive (18) than negative items (15). Selecting almost exclusively negative items for the short form might therefore decrease the likelihood that the short form measures the same construct as the original scale. While belonging to the same construct, negative and positive items tap into slightly different aspects of SDRS i.e. that of not admitting undesirable behavior (something that one has done) and pretending to behave in a desirable way (something that one has not done). Focusing mainly on one aspect may thus create a more narrow measure of SDRS. To select mainly negative items might also increase the response burden of the scale, because all of the negative items are categorized under at least one of the non-clarity categories and slightly more often than the positive items under two out of the three categories. Negatively worded items thus seem to be more burdensome for respondents, a well-known finding in the survey methodology literature (see e.g. Lyberg et al., 1997).

Another thing to consider are the *b* estimates of the remaining items. All items with *b* estimates falling outside of the typical range (±2) were deleted based on low *a* estimates, except item 20 (*When I don't know something I don't at all mind admitting it*). The *b* estimates for the other 23 items range between −1.84 and 1.48 (item 9 and item 5, respectively), which makes item 20 the by far most sensitive item with a *b* estimate of −2.41 (endorsed by 87.3% of the sample in the neutral setting, in which the item was presented in the current study). Such a sensitive item might not be well suited for the purpose of creating a short form of the MCSDS to detect SDRS in Internet surveys; in cases where there is reason to believe that the responses are distorted by SDRS because the sample taking the survey is likely to present themselves favorably (either because of the situation in which the survey is taken or because of characteristics of the sample, or both), such an item would presumably be endorsed by such a large proportion of the sample that it would cease to be informative (much like an elementary school math problem would not be very useful in scheming for high school level math proficiency). Thus, item 20 was not included in further analysis. However, this does not mean that item 20 is not a good item for the purpose of obtaining accurate estimates of SDRS as it has a relatively high discrimination estimate and does not fall under any non-clarity category.

Items with very low and low estimates of *a* and items with low estimates of *b* were excluded based on these criteria. By using these criteria, the items that repeatedly obtained low factor loadings (i.e. in all three studies reported in Table 3) were all dropped (item 7, 18 and 29). A total of ten items were eliminated from the MCSDS: Items 1, 2, 7, 8, 17, 18, 20, 24, 25 and 29.

3.4. Item selection: suggested short form

The item elimination procedure leaves 23 items. The MCSDS is used with other measures (to validate them and/or control for the effect of SDRS) and is therefore always an addition to surveys consisting of at least one other scale. Adding 23 items to a survey would add substantial respondent burden, thus increasing the risk of response biases due to fatigue. For this reason the list of remaining items was further reduced. For the short form, items were selected *in* instead of *out* as in the item elimination process.

3.4.1. Positive items

Out of the remaining positive items, items 4, 13, 16, 31 and 33 achieved the highest discrimination estimates. These items all have overall moderate discrimination estimates and are not located at the same sensitivity level, nor were their factor loadings below 0.30 in any of the three studies cited in Table 3.

3.4.2. Negative items

The six most discriminating items are all negatively worded, the best five of these were selected for the short form: items 6, 15, 19, 23 and 28. The sixth most discriminating item, item 9, was excluded due to the item's low sensitivity estimate: −1.84, a factor loading below 0.30 in one study and a classification under two of the five non-clarity categories. Items 14 and 30 were also considered for the short form on the basis of their discrimination estimates. These items were however excluded because they have similar sensitivity estimates to items that have already been included in the short form (items 33 and 15, respectively).

3.5. Psychometric properties of the short form

The short form was made up of the following items: 4, 6, 13, 15, 16, 19, 23, 28, 31 and 33 (see Table 5). Results from CFA analysis of the short form are presented in Table 4. The fit of a one factor model was above the cutoff criterion for an adequate fit in terms of χ^2/df and CFI, TLI, SRMR and RMSEA indicated a good fit. The Chi-square test was however significant.

The composite reliability for the MCSD-SF was good ($\rho x = 0.83$) but a bit lower than for the MCSDS ($\rho x = 0.89$).

IRT analysis was run for the MCSD-SF. Items and item parameters are displayed in Table 5. The item sensitivity estimates of the MCSD-SF have a rather limited range (−1.03 to 1.15). This restricts measurements obtained with the MCSD-SF to the extent that the accuracy of estimates is reduced for those with high and low trait levels. As low levels of SDRS are not expected to cause problems (i.e. result in distortion of other measurements due to favorable self presentation), accuracy at lower levels is of less importance. On the other hand, it would have been desirable to include more items with higher estimates of *b*. However, the items with thresholds at the upper end of the score distribution (items that are less sensitive to SDRS) have rather low discrimination estimates. Selecting the items that are less sensitive to SDRS would thus have been at the expense of item discrimination.

The item characteristic curves (ICC) and the item information functions (IIF) are shown in Fig. 2, and Fig. 3 shows the test information functions (TIF), for the MCSDS and the MCSD-SF.

As can be expected, some information is lost when reducing the scale from 33 items to only ten. The practical benefits of administering only ten items instead of 33 should however outweigh the drawbacks.

4. Discussion

The goal of this study was to eliminate items with the weakest psychometric properties and develop a psychometrically sound short form of the MCSDS, using a combination of IRT analysis of Internet panel data and CogI, along with previously obtained CFA results. First, nine items

Table 4
Confirmatory factor analysis of the MCSD-SF.

	N of items	χ^2	*p*-Value	*df*	χ^2/df	CFI	TLI	SRMR	RMSEA	RMSEA 90% CI
MCSD-SF	10	52.93	0.026	35	1.51	0.99	0.98	0.059	0.031	0.011; 0.047

were excluded because of poor psychometric properties and the 10th item (item 20) was dropped because of its potential over-sensitivity in non-neutral situations. These items can thus be eliminated from further examinations of the scale, with the possible exception of item 17. The poor psychometric properties of this item are likely due to the difficulty of translating a phrase to a language that does not have a corresponding phrase. Item 17 may therefore not prove to be problematic if used in the scale's original language (English) or in translated versions that capture the item's content sufficiently.

Without further research it is impossible to know, how and to what extent using a translated version affected the results and whether the results obtained here generalize to other language versions of the scale. It would therefore be informative to replicate this study in other languages and compare the results. However, incompatibility between translations of an item is not always easily dealt with, as the change in properties between translations might be caused by something other than wording (e.g. unclarity of the item, culture etc.). In any case it should be safe to assume that the clearer the wording and meaning of the original item, the easier it is to translate it to another language. It is important that a scale that is used internationally contains clear and translatable items. Therefore poor psychometric properties obtained for any item in any language (given that the research was well conducted) should always be of concern to researchers, although it cannot be taken for granted that the findings generalize and item elimination should always be done with caution.

Initial elimination of items resulted in a 23 items scale. However, since MCSDS is always an addition to other measures, adding a scale this long could still be overly costly and increase the risk of response bias. For these reasons the list of remaining items was further reduced and ten items were selected for the short form. It must be pointed out that the results of this study do not provide reasons for dropping these 13 items altogether. It is therefore recommended that these items are included in future research on the on the MCSDS. However, if the purpose of a study is to use the MCSDS to assess or control for systematic measurement error due to SDRS and cost or response burden is of concern, using our ten item short form instead of 23 items may be beneficial.

The ten item short form produced in this study has good psychometric properties and reduces response burden substantially and is therefore less likely to exhaust respondents' patience and reduce response quality (Galesic & Bosnjak, 2009). We generally recommend using a one-item-one-screen (OIOS) format in Internet-based questionnaires (Reips, 2010). However, ten items also fit easily on a normal computer screen, which makes scrolling or page flipping unnecessary to respond to the MCSD-SF on a normal sized computer and would reduce scrolling or page flipping on smaller devices used for Internet survey participation compared to the original MCSDS. The MCSD-SF also contains both positively and negatively worded items enabling the detection of response biases such as straight lining (responding to all items in the same response category). However, the psychometric properties of the MCSD-SF are calculated from responses to the full 33 item scale, which ignores the possibility that response patterns may change if these items are presented sequentially (due to different context effects). The items selected for the MCSD-SF would therefore have to be administered as a short form and its psychometric properties need to be reevaluated.

The MCSD-SF was developed for Internet administration and social desirable responding has been shown to work differently on the Internet (Kaufmann & Reips, 2008). As discussed by Tourangeau et al. (2000), mode of data collection can affect responses to survey questions and therefore results obtained in one mode may not generalize to other modes. This may be of special concern when measuring social desirability because studies have generally found that self-administrated measures are less affected by SDRS than interviewer administrated measures (see for example Kreuter et al., 2008). The results become less clear when online modes are compared to paper and pencil administrations. In a meta-analysis Gnambs and Kaspar (2015) found that computerized surveys led to more reporting of undesirable behavior than paper and pencil surveys (especially when the behavior was highly sensitive). However, meta-analyses of computerized and paper and pencil administrations of social desirability scales and personality measures (Dodou & de Winter, 2014; Gnambs & Kaspar, 2016) found no overall difference in social desirability between the two self-administration modes. When looking specifically at the MCSDS, the meta-analysis conducted by Dodou and de Winter (2014) found that higher values were obtained with the MCSDS in paper and pencil than computerized administrations, whereas Gnambs and Kaspar (2016) reported no mode effects.

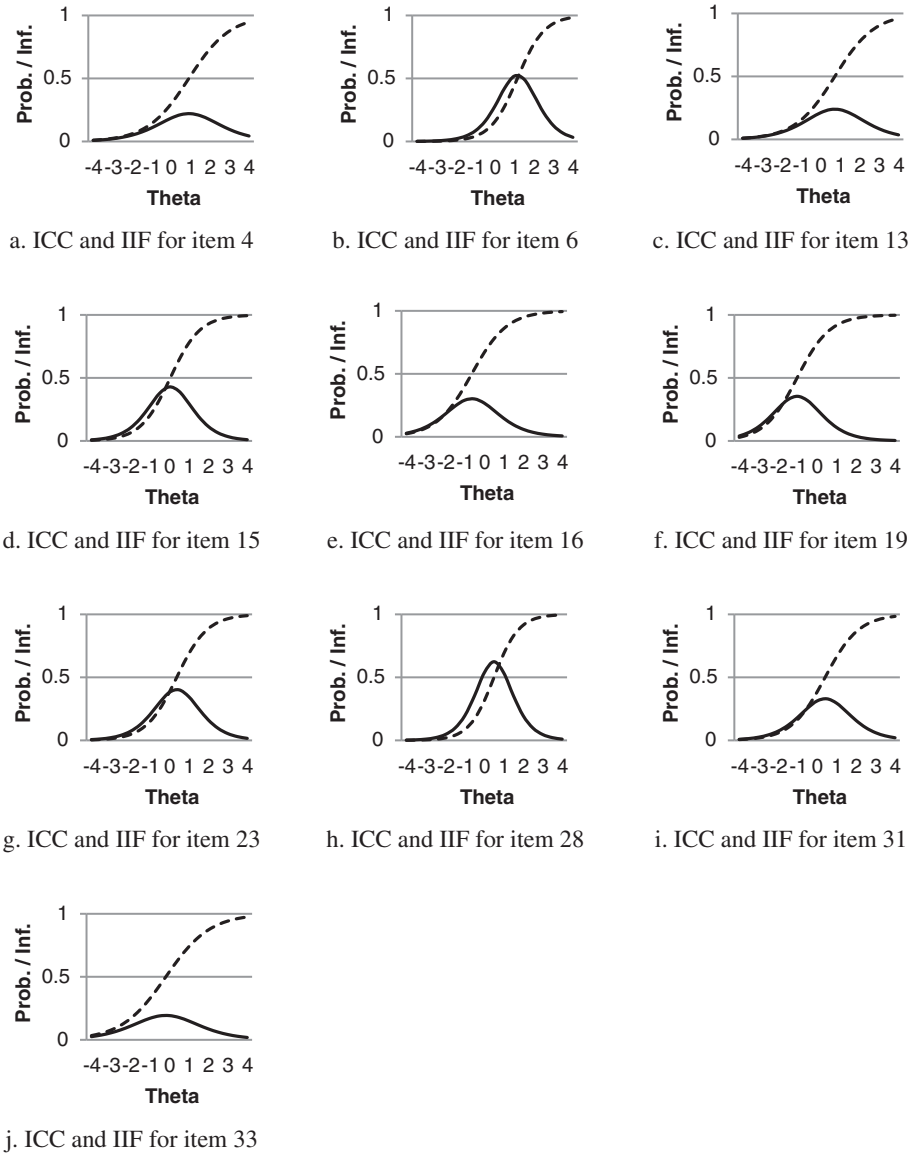
It is thus not entirely clear how self-administered modes affect socially desirable responding. Furthermore, psychometric properties can never be assumed to be stable between modes, nor can mode effects be presumed to be stable over time. Concerns over Internet privacy may for example increase and thus online studies may become more susceptible to the effects of SDRS than other forms of data collection. The bottom line is, not to think of psychometric properties as properties of the scale (although we often speak of properties in that manner) but as properties of measurements obtained with the scale in a certain setting, and always be cautious when generalizing.

The concept measured by the MCSD-SF seems to describe more general rules of conduct than the MCSDS, as all the items that fall under the non-clarity category *Presumptions*, i.e. items that specify certain things or specific situations (items 1, 8, 9 and 27) have been removed. This

Table 5
IRT parameter estimates of the short form items.

Item		*a*	*b*
4.	I have never intensely disliked anyone	0.94 (0.15)	0.91 (0.16)
*6.	I sometimes feel resentful when I don't get my way	1.45 (0.21)	1.15 (0.13)
13.	No matter who I'm talking to, I'm always a good listener	0.98 (0.15)	0.73 (0.14)
*15.	There have been occasions when I took advantage of someone	1.31 (0.18)	0.05 (0.09)
16.	I'm always willing to admit it when I make a mistake	1.10 (0.17)	−0.64 (0.12)
*19.	I sometimes try to get even, rather than forgive and forget	1.19 (0.18)	−1.03 (0.15)
*23.	There have been occasions when I felt like smashing things	1.27 (0.18)	0.38 (0.10)
*28.	There have been times when I was quite jealous of the good fortune of others	1.58 (0.22)	0.50 (0.09)
31.	I have never felt that I was punished without cause	1.15 (0.16)	0.41 (0.10)
33.	I have never deliberately said something that hurt someone's feelings	0.88 (0.14)	−0.18 (0.12)

Note: a = discrimination parameters, b = sensitivity parameters. Standard errors are shown in brackets. Items marked with * are keyed in the false direction.



a. ICC and IIF for item 4 b. ICC and IIF for item 6 c. ICC and IIF for item 13

d. ICC and IIF for item 15 e. ICC and IIF for item 16 f. ICC and IIF for item 19

g. ICC and IIF for item 23 h. ICC and IIF for item 28 i. ICC and IIF for item 31

j. ICC and IIF for item 33

Fig. 2. Item characteristic curves (dash line) and item information functions (solid line) for short form items.

may be because when an item specifies certain things or circumstances, those things or circumstances may not fit to the respondent's experience and therefore the respondent tries to fit the item by reinterpreting it. So, to a certain extent, the more specific the item content the more varied the interpretation. It is also more likely that such items are culture and time specific. With one exception (item 23), the MCSD-SF items describe human interactions or the respondents' reaction to such interactions, which is something that the vast majority of people can relate to. The MCSD-SF may therefore be more applicable to more diverse groups than the MCSDS.

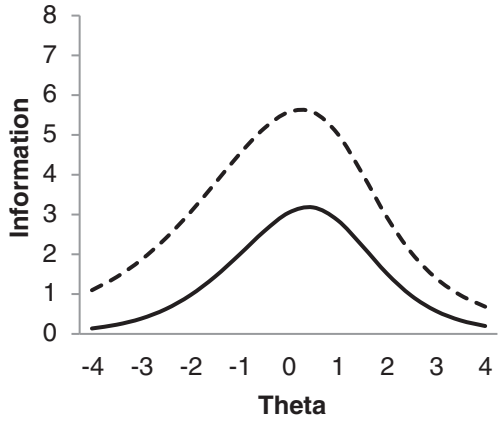


Fig. 3. Test information functions for the full 33 item MCSDS (dash line) and the 10 item MCSD-SF (solid line).

Combining IRT and CogI seems to have been an effective way to evaluate items and develop a short form, and this procedure is certainly superior to focusing solely on internal consistency. However, combining these two methods has its limitations. The MCSD-SF is intended for Internet use but the CogI were conducted face to face. The information obtained in the CogI is thus obtained in a very different setting from that in which the panel data is collected. It must also be kept in mind as a limitation that the CogI are based on 40 individuals and no cutoff criterion was set for how many of them had to find something confusing about an item for the item to fall under a non-clarity category. Furthermore, it is possible that this method of conducting CogI (having participants answer probing questions after each item) produces an overly sensitive measure of clarity, because the interviewees are primed to think very

carefully about the clarity of each item (which is something that a normal Internet survey respondent cannot be expected to do). Thus, the possibility, that interviewees do not only base their responses on how they initially understood the item but also on how the item *could* be understood, should be taken into account. CogI can provide very valuable information that cannot be obtained by other means, but it is important to acknowledge the limitations of this method and be cautious when interpreting the results.

Another issue to consider is that, due to the nature of the construct, it may not be possible to clear up all the non-clarity indicators identified in the CogI. The MCSDS is made up of items that describe "behaviors which are culturally sanctioned and approved but which are improbable of occurrence" (Crowne & Marlowe, 1960, p. 350). It is assumed that respondents are exaggerating when giving a response in the keyed direction. Frequency is thus an essential part of the construct, so although people may have a different understanding of words and phrases that describe frequency, references to the frequency of behavior cannot be removed from the items without changing the construct being measured. It would be informative though to have some estimates of the actual frequency of the behaviors in question. This would make it possible to estimate the likelihood of responding honestly to all ten items in the keyed direction, and the increase/decrease in that likelihood with each item dropped/added to the scale. With each item that is dropped from the scale the chances of responding honestly to all items in the keyed direction increase. If, for example, a common undesirable behavior is performed by 70% of people there is a 30% chance that the respondent does not behave in that way and could honestly say he/she does not. If we add another item with the same base rate frequency the chances decrease and become approximately $0.30 * 0.30$ ("approximately" because it is not unlikely that someone belonging to the 30% on the first item will have a greater chance of belonging to the 30% on the second item).

Given the above limitations, it may prove useful to use other techniques in the future, e.g. latent class analysis (e.g. Dantlgraber et al., 2016). The ultimate test is of course whether the items prove to be useful, i.e. whether SDRS in Internet surveys can be dealt with by controlling for responses to the MCSD-SF. Further research is needed to answer this question. However, shortening the MCSDS, removing items with poor psychometric properties and refining the construct to focus more on human feelings and interactions, increases the probability that this can be accomplished. After all, the mere assumption of socially desirable responding - the systematic tendency to respond to survey items in a manner that will be *viewed favorably by others* - should lead to the assumption that items on one's *behavior and feeling towards others* should be affected by this tendency.

Acknowledgements

This research was partly funded by The Eimskip Fund of The University of Iceland (Háskólasjóður Eimskipafélags Íslands). The funding source had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; nor in the decision to submit the article for publication.

The authors would like to acknowledge networking support by the COST Action IS1004. www.webdatanet.eu

The authors would like to thank Bylgja Björk Pálsdóttir and Soffía Svanhildar Felixdóttir for conducting and assisting with the analysis of the cognitive interviews used in this study. The authors would also like to thank Christopher Desjardins for helpful comments on the paper.

References

Baker, F. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Ballard, R. (1992). Short forms of the Marlowe-Crowne social desirability scale. *Psychological Reports*, *71*, 1155–1160. http://dx.doi.org/10.2466/PR0.71.8.1155-1160.

Barger, S. D. (2002). The Marlowe-Crowne affair: Short forms, psychometric structure, and social desirability. *Journal of Personality Assessment*, *79*(2), 286–305. http://dx.doi.org/10.1207/S15327752JPA7902_11.

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne social desirability scale. *Educational and Psychological Measurement*, *62*, 570–589. http://dx.doi.org/10.1177/0013164402062004003.

Black, J. E., & Reynolds, W. M. (2016). Development, reliability, and validity of the moral identity questionnaire. *Personality and Individual Differences*, *97*, 120–129. http://dx.doi.org/10.1016/j.paid.2016.03.041.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319. http://dx.doi.org/10.1037/1040-3590.7.3.309.

Cote, J. A., & Buckley, M. R. (1988). Measurement error and theory testing in consumer research: An illustration of the importance of construct validation. *Journal of Consumer Research*, *14*, 579–582. http://dx.doi.org/10.1086/209137.

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

Dantlgraber, M., Wetzel, E., Schützenberger, P., Stieger, S., & Reips, U. -D. (2016). Simple construct evaluation with latent class analysis: An investigation of Facebook addiction and the development of a short form of the Facebook Addiction Test (F-AT). *Behavior Research Methods*, *48*, 869–879. http://dx.doi.org/10.3758/s13428-016-0716-2.

Davis, C. G., Thake, J., & Weekes, J. R. (2012). Impression managers: Nice guys or serious criminals? *Journal of Research in Personality*, *46*, 26–31. http://dx.doi.org/10.1016/j.jrp.2011.11.001.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Los Angeles: Sage.

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. http://dx.doi.org/10.1016/j.chb.2014.04.0053.

Edelen, M. O., & Reeve, B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5–18. http://dx.doi.org/10.1007/s11136-007-9198-0.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Fischer, D. G., & Fick, C. (1993). Measuring social desirability: Short forms of the Marlowe-Crowne social desirability scale. *Educational and Psychological Measurement*, *53*, 417–424. http://dx.doi.org/10.1177/0013164493053002011.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage Publications.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. http://dx.doi.org/10.1093/poq/nfp031.

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

Gnambs, T., & Kaspar, K. (2016). Socially desirable responding in web-based questionnaires: A meta-analytic review of the candor hypothesis. *Assessment*. http://dx.doi.org/10.1177/1073191115624547 (Advance online publication).

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory (measurement methods for the social sciences series, Vol. 2)*. Newbury Park, CA: Sage Publications.

Helmes, E., & Holden, R. R. (2003). The construct of social desirability: One or two dimensions? *Personality and Individual Differences*, *34*, 1015–1023. http://dx.doi.org/10.1016/S0191-8869(02)00086-7.

Hu, L. -T., & Bentler, P. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. http://dx.doi.org/10.1080/10705519909540118.

Kang, T., Cohen, A. S., & Sung, H. J. (2005). IRT model selection methods for polytomous items. *Annual meeting of the National Council on Measurement in Education, Montreal*, March.

Kaufmann, E., & Reips, U. -D. (2008). *Internet-basierte Messung sozialer Erwünschtheit [Internet-based measurement of social desirability]*. Saarbrücken: VDM Verlag Dr. Müller.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*, 847–865. http://dx.doi.org/10.1093/poq/nfn063.

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

Leite, W. L., & Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne social desirability scale and the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *65*, 140–154. http://dx.doi.org/10.1177/0013164404267285.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. http://dx.doi.org/10.2307/2290157.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*, 493–504. http://dx.doi.org/10.1037/h0058543.

Loo, R., & Thorpe, K. (2000). Confirmatory factor analyses of the full and short versions of the Marlowe-Crowne social desirability scale. *The Journal of Social Psychology*, *140*, 628–635. http://dx.doi.org/10.1080/00224540009600503.

Lyberg, L., Biemer, P., Collins, M., deLeeuw, E., Dippo, C., Schwarz, N., & Trewin, D. (Eds.). (1997). *Survey measurement and process quality*. New York, NY: Wiley.

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, *51*(6), 882–888. http://dx.doi.org/10.1037/0022-006X.51.6.882.

Pace, V. L. (2010). Method variance from the perspectives of reviewers: Poorly understood problem or overemphasized complaint. *Organizational Research Methods*, *13*(3), 421–434. http://dx.doi.org/10.1177/1094428109351751.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. http://dx.doi.org/10.1037/0022-3514.46.3.598.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Podsakoff, P. M., MacKenzie, S. M., Podsakoff, N. P., & Lee, J. (2003). The mismeasure of man(agement) and its implications for leadership research. *The Leadership Quarterly*, *14*, 615–656. http://dx.doi.org/10.1016/j.leaqua.2003.08.002.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *65*, 539–569. http://dx.doi.org/10.1146/annurev-psych-120710-100452.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173–184. http://dx.doi.org/10.1177/01466216970212006.

Reips, U. D. (2010). Design and formatting in Internet-based research. In S. Gosling, & J. Johnson (Eds.), *Advanced methods for conducting online behavioral research*. (pp. 29–43). Washington, DC: American Psychological Association.

Reips, U. -D. (2012). Using the Internet to collect data. In H. Cooper, P. M. Camic, R. Gonzalez, D. L. Long, A. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology. Research designs: Quantitative, qualitative, neuropsychological, and biological*, *2*. (pp. 291–310). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/13620-017.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*(2), 93–103. http://dx.doi.org/10.1207/S15327752JPA8102_01.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology*, *38*, 119–125. http://dx.doi.org/10.1002/1097-4679(198201)38:1%3C119::AID-JCLP2270380118%3E3.0.CO;2-I.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25 (http://www.jstatsoft.org/v17/i05/).

Rosseel, Y. (2012). *lavaan*: An R *package* for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://dx.doi.org/10.18637/jss.v048.i02.

Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Beverly Hills, CA: Sage Publications.

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*(1), 167–193. http://dx.doi.org/10.1111/j.1744-6570.2002.tb00108.x.

Stark, S. (2001). *MODFIT: A computer program for model-data fit. Unpublished manuscript.* University of Illinois: Urbana-Champaign.

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlowe-Crowne social desirability scale. *Journal of Clinical Psychology*, *28*, 191–193. http://dx.doi.org/10.1002/1097-4679(197204)28:2<191::AID-JCLP2270280220>3.0.CO;2-G.

Tay, L., & Drasgow, F. (2012). Adjusting the adjusted χ^2/df ratio statistic for dichotomous item response theory analyses: Does the model fit? *Educational and Psychological Measurement*, *72*(3), 510–528. http://dx.doi.org/10.1177/0013164411416976.

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

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Tracey, T. J. G. (2016). A note on socially desirable responding. *Journal of Counseling Psychology*, *63*(2), 224–232. http://dx.doi.org/10.1037/cou0000135.

van Schie, K., Wanmaker, S., Yocarini, I., & Bouwmeester, S. (2016). Psychometric qualities of the thought suppression inventory-revised in different age groups. *Personality and Individual Differences*, *91*, 89–97. http://dx.doi.org/10.1016/j.paid.2015.11.060.

Ventimiglia, M., & MacDonald, D. A. (2012). An examination of the factorial dimensionality of the Marlowe Crowne social desirability scale. *Personality and Individual Differences*, *52*, 487–491. http://dx.doi.org/10.1016/j.paid.2011.11.016.

Vésteinsdóttir, V., Reips, U. -D., Joinson, J., & Thorsdottir, F. (2015). Psychometric properties of measurements obtained with the Marlowe–Crowne social desirability scale in an Icelandic probability based Internet sample. *Computers in Human Behavior*, *49*, 608–614. http://dx.doi.org/10.1016/j.chb.2015.03.044.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* (Doctoral dissertation) Los Angeles: University of California (2002).