Behav Res (2013) 45:527–535
DOI 10.3758/s13428-012-0265-2

Seriousness checks are useful to improve data validity in online research

Frederik Aust · Birk Diedenhofen · Sebastian Ullrich · Jochen Musch

Published online: 10 October 2012

Abstract Nonserious answering behavior increases noise and reduces experimental power; it is therefore one of the most important threats to the validity of online research. A simple way to address the problem is to ask respondents about the seriousness of their participation and to exclude self-declared nonserious participants from analysis. To validate this approach, a survey was conducted in the week prior to the German 2009 federal election to the Bundestag. Serious participants answered a number of attitudinal and behavioral questions in a more consistent and predictively valid manner than did nonserious participants. We therefore recommend routinely employing seriousness checks in online surveys to improve data validity.

Keywords Online research · Participant screening · Seriousness · Voting survey · Incremental validity · Methodology

The Internet provides an attractive environment for the convenient large-scale collection of data (Couper, 2000; Fricker & Schonlau, 2002; Reips, 2000, 2011). Moreover, collecting data online provides an opportunity to conduct surveys targeting otherwise difficult-to-reach populations (Mangan & Reips, 2007; Reips & Buffardi, 2012). The perceived anonymity of Web-based surveys may also reduce socially desirable answering behavior and help to obtain more reliable self-reports of sensitive behavior (Davis, 1999; Joinson, 1999; King & Miles, 1995; Reips, 2011). For example, in a large survey of American adolescents, Turner et al. (1998) found that reports of socially stigmatizing behavior in a self-administered questionnaire were greatly increased when computers were being used to collect responses. Online surveys have since increasingly been used to address substantive research problems.

A problem that comes with the easy accessibility of online studies, however, is the large diversity of participants taking part in Web surveys. People just browsing the Internet for interesting content may provide useless data when clicking through a questionnaire out of curiosity, rather than providing well thought out answers (Reips, 2009). Moreover, researchers or visitors solely interested in having a look at a study's methodology and research question are frequently forced to submit data even if they are not motivated to provide valid responses (Reips, 2009). These are problems concerning all types of publicly accessible Internet-based research. The main problem resulting from the participation of nonserious respondents, however, is that they increase noise and reduce experimental power. This poses a serious threat to the validity of online research (Oppenheimer, Meyvis, & Davidenko, 2009; Reips, 2002, 2009).

Consistency checks

One way to address the problem is to analyze the consistency and plausibility of the answers in an attempt to identify participants who did not answer seriously (e.g., Reips, 2000, 2002, 2009). For example, when demographic data are

F.A. and B.D. contributed equally to this study; authorship order was determined by a coin.

F. Aust (✉) · B. Diedenhofen (✉) · S. Ullrich (✉) · J. Musch (✉)
Department of Experimental Psychology,
University of Duesseldorf,
Universitaetsstrasse 1, Building 23.03,
40225 Duesseldorf, Germany
e-mail: frederik.aust@hhu.de
e-mail: birk.diedenhofen@hhu.de
e-mail: sebastian.ullrich@hhu.de
e-mail: jochen.musch@hhu.de

collected, checking for implausible or impossible combinations of age and education level or income may reveal low-quality data sets. Unfortunately, however, this approach to reducing noise can be employed only if impossible combinations of answers can be identified and can be screened for without falsely identifying serious submissions as invalid.

Unique IP check

Another approach is to restrict analyses to data sets with unique IP addresses to alleviate biasing effects of multiple submissions. This is considered to be a conservative procedure because multiple users may share the same IP address—for example, when accessing the survey via a proxy server (Reips, 2000).

Completion time check

Excluding participants with exceedingly short completion times is another measure that has been used to reduce noise (e.g., Ihme et al., 2009; Keller, Gunasekharan, Mayo, & Corley, 2009; Lahl, Göritz, Pietrowsky, & Rosenberg, 2009; Malhotra, 2008; Musch & Klauer, 2002). Speeders may save time by skimming over instructions, performing shallow memory searches, making hasty judgments, or simply answering randomly. However, appropriate thresholds necessary to identify speeding are difficult to determine and depend on the distribution of response times (Ratcliff, 1993). Alternative indirect approaches to testing the carefulness with which respondents submitted their data are therefore of interest.

Instructional manipulation check

One approach, proposed by Oppenheimer et al. (2009), tries to identify respondents who did not even read the instructions,

> a question embedded within the experimental material that is similar to the other questions in length and response format (e.g., Likert scale, check boxes, etc.). However, unlike the other questions, the [Instructional Manipulation Check] asks participants to ignore the standard response format and instead provide a confirmation that they have read the instructions. (p. 867)

Oppenheimer et al. found evidence that the exclusion of participants failing the instructional manipulation check improved data quality. The size of the effects that could be observed in two classic judgment and decision-making paradigms was larger, and a Need for Cognition scale was answered in a more consistent manner by participants passing the check. An approach that was also shown to be effective by Oppenheimer and colleagues is to prompt those failing the check to carefully reread the instructions. After such a prompt, the answer behavior of participants initially failing the check became indistinguishable from that of participants passing the check (Oppenheimer et al., 2009), thus improving data quality without losing data and risking selection bias. The applicability of this innovative procedure is, however, limited to studies in which data quality is dependent on the careful reading of instructions. In such cases, the prompt to ignore the standard response format can be embedded in the question text. However, it cannot be presumed that skipping instructions reduces the quality of answers unless they provide essential information. With the use of an instructional manipulation check, there might also be a risk of participants foiling the study because they take offense, feeling distrusted or embarrassed upon failing the check.

Person fit indices

Rasch person fit indices offer a methodologically advanced approach to detecting aberrant responses (Li & Olejnik, 1997). They can be used, for example, to identify atypical response patterns that may occur as a result of cheating (Madsen, 1987) or socially desirable answering behavior (Schmitt, Cortina, & Whitney, 1993). However, the applicability of such tests is limited. They can be employed only if a test was constructed to fit an item response model (Li & Olejnik, 1997; van den Wittenboer, Hox, & de Leeuw, 1997), and their power highly depends on the length of a test and the range of the difficulties of its items (Reise & Due, 1991).

Seriousness check

A very economic measure for identifying nonserious participants was used by Musch and Klauer (2002; cf. Reips, 2002). They directly asked the respondents to indicate the seriousness of their responses. Nonserious participants were thus able to identify themselves. Some studies already indicated the usefulness of this procedure. For example, the seriousness of a participation reported at the beginning of a study has been shown to be the best predictor of dropout rates and, thus, a measure of motivation (Reips, 2002, 2008, 2009). However, a direct investigation of data quality improvement has not been reported yet. In an early application of the technique, Klauer Musch, and Naumer (2000) used the seriousness check in combination with the exclusion of multiple submissions on the basis of the respondents' IP and e-mail addresses and the deletion of trials with response times outside a defined temporal window. Although the analyses provided results that converged with data collected

in the laboratory, it was not tested whether the deletion of responses affected the findings. Accordingly, Klauer et al. provided no validation of the seriousness check.

Although asking respondents directly about the seriousness of their participation may seem an obvious improvement, an analysis of articles published between 2009 and 2010 in three major journals reporting online studies (*Behavior Research Methods*, *International Journal of Internet Science*, and *International Journal of Human–Computer Studies*) suggests that seriousness checks are used rarely (Table 1). Out of 32 studies recruiting participants online, only 6 (18.8 %) reported the use of one or more measures to ensure or improve data quality. In three cases (9.4 %), multiple participations were detected by logging the respondents' IP addresses. Four studies (12.5 %) checked for inconsistent answers, and three studies (9.4 %) considered the survey completion time. Egermann et al. (2009) was the only study (3.1 %) employing a seriousness check. Pursuing a similar approach, Buchanan et al. (2010) asked their respondents whether there were reasons their data should not be used, after having answered a questionnaire on executive function. In a similar vein, Ihme et al. (2009) administered ability tests online and assessed their participants' concentration by directly questioning them as an indicator for data quality. In spite of these occasional uses, the utility and validity of employing seriousness checks have never been scrutinized. Instead, their use seems to have primarily been founded on plausibility, and no analysis comparing data with and without participants differing regarding the seriousness of their responses has been reported. When a seriousness check or a similar measure was used, the rate of self-reported nonserious participations ranged from 5 % – 6 % (Musch & Klauer, 2002) to 30 % – 50 % Reips (2009). These figures suggest that it might be important to conduct a more thorough analysis of the problem of nonserious participants and the effects their participation might have on the quality of data collected online. For these reasons, we decided to conduct a first thorough investigation of the seriousness check.

The primary goal of the present study was to investigate the extent to which the data of self-reported nonserious participants differ from serious submissions and whether their exclusion has the potential to increase the validity of data collected online. Our major hypothesis was that serious participants would provide more coherent and valid data than would nonserious participants. We therefore expected that in an online survey, conducted in the week prior to the German 2009 federal election to the Bundestag, serious participants would report more consistent combinations of political attitudes and sympathies toward potential government formations. Similarly, given that individual voting behavior is at least partially stable (Schmitt, Sanz, & Braun, 2009), we expected the accordance of voting intention for 2009 and voting behavior in 2005 to be higher for serious participants not distorting their self-reported intentions and the report of their past behavior. Finally, we also expected the forecast of the final election results to be more accurate if it was exclusively based on the answers of serious respondents. Our investigation also allowed us to explore the incremental benefit of seriousness checks when combined with other common methods of participant screening—namely, a unique IP check, a completion time check, and a consistency check.

Method

A total of 3,786 German participants were recruited for the online survey through a Google AdWords campaign using election-related keywords on Google's search result pages. Participants who were not entitled to vote because of their age or for other reasons and respondents who refused to report their voting intentions were excluded from analysis. This resulted in a final data set consisting of 3,490 respondents. Of this final sample of respondents, 76.0 % were male, 62.6 % had at least 11 years of formal education or had an academic degree, and 73.4 % reported being between 18 and 44 years

Table 1 Methods employed to ensure data quality in 32 online studies published in three journals between 2009 and 2010

	Checks			
Article	Unique IP	Consistency	Completion time	Seriousness
Buchanan et al., 2010	•	•		•
Egermann, Nagel, Altenmüller, & Kopiez, 2009				•
Ihme et al., 2009	•		•	•
Keller, Gunasekharan, Mayo, & Corley, 2009	•		•	
Lahl, Göritz, Pietrowsky, & Rosenberg, 2009	•	•	•	
Whitty & Buchanan, 2010		•		

Note. Two related methods were also scored as a seriousness check: Buchanan et al. asked their respondents whether there were reasons their data should not be used. Ihme et al. assessed participants' concentration by direct questioning. Whereas Egermann et al. employed their seriousness check at the beginning of their study, Buchanan et al. and Ihme et al. collected their respective measures after the completion of their studies. No study assessed the beneficial effect of this seriousness check, however.

old. The sample was thus subject to the well-documented bias toward young and well-educated participants researchers frequently face when conducting studies on the Web (Krantz & Dalal, 2000; Schmidt, 1997).

On the first two pages, participants provided demographic information in accordance with the German microcensus, an official representative statistic of the German population conducted by the German Federal Statistical Office. On each of the following pages, respondents reported their voting intention for the upcoming election and the vote they had cast in the previous federal election in 2005. Participants also reported their political attitudes on a left–right scale and their sympathy toward a possible government participation of the five major parties (Conservatives, Social Democrats, Liberals, Left, and Greens) and two major potential coalitions (Conservatives + Liberals and Social Democrats + Left + Greens).

The seriousness of participation was assessed separately on a single page at the end of the survey. Participants were asked whether they had answered in a serious manner, so that their responses could be used to investigate research questions on political attitudes. The wording of the question was the following: "It would be very helpful if you could tell us at this point whether you have taken part seriously, so that we can use your answers for our scientific analysis, or whether you were just clicking through to take a look at the survey?" Participants were able to choose one of two answers: "I have taken part seriously" or "I have just clicked through, please throw my data away." To investigate the methodological research question at hand, we analyzed self-declared nonserious data sets and present the results in summarized form.

Results

The probability of a type I error was set at .05 for all subsequent tests. Of the 3,490 valid participants, 3,378 (96.8 %) reported having answered seriously, whereas 112 (3.2 %) identified their responses as nonserious. Previous research has shown that the Social Democrats, the Left, the Green party, and Social Democrats + Left + Greens coalition are usually regarded to be left-wing oriented, whereas the Conservative party, the Liberals, and a potential Conservatives + Liberals coalition are considered to be more right-wing by German voters (Pappi, 2009). As a first index of data validity, we therefore calculated correlations between the participants' self-reported political attitudes on a left–right scale and their self-reported sympathies toward a future government participation of the five major parties and two potential coalitions.

In one-tailed z tests conducted according to formula 2.8.11 in Cohen, Cohen, West, and Aiken (2003, p. 49), serious participants' self-reported political attitudes on a left–right scale correlated significantly higher with their sympathies toward a 2009 government participation (Table 2). This was true for the Social Democratic Party, $r_{serious} = -.37$, $r_{nonserious} = -.19$, $z = -2.04$, $p < .03$, the Liberal party, $r_{serious} = .51$, $r_{nonserious} = .23$, $z = 3.39$, $p < .01$, and the Conservatives + Liberals coalition, $r_{serious} = .62$, $r_{nonserious} = .42$, $z = 2.81$, $p < .01$. For the Conservative party, $r_{serious} = .57$, $r_{nonserious} = .45$, $z = 1.64$, $p < .06$, the Green party, $r_{serious} = -.41$, $r_{nonserious} = -.29$, $z = -1.39$, $p < .09$, and the Social Democrats + Left + Greens coalition, $r_{serious} = -.58$, $r_{nonserious} = -.48$, $z = -1.46$, $p < .08$, the effect of seriousness was also—albeit, marginally—significant. For the Left party, correlations differed only descriptively, $r_{serious} = -.53$, $r_{nonserious} = -.44$, $z = -1.25$, $p < .11$. Thus, taken together, serious participants answered attitudinal questions in a more consistent manner than did nonserious participants.

Table 2 Different correlations of serious and nonserious respondents' self-reported left–right classification and their sympathy toward governmental participation of the major parties and coalitions for the German 2009 federal election

	Participants			
Parties and Coalitions	Serious	Nonserious	z	p
Conservatives	.57	.45	1.64	< .06
Social Democrats	−.37	−.19	−2.04	< .03
Liberals	.51	.23	3.39	< .01
Greens	−.41	−.29	−1.39	< .09
Left	−.53	−.44	−1.25	< .11
Conservatives + Liberals	.62	.42	2.81	< .01
Social Democrats + Left + Greens	−.58	−.48	−1.46	< .08

Note. In every case, except for the Left, the magnitude of the correlation was significantly or marginally significantly larger for serious participants (one-tailed tests).

Further evidence for the utility of asking about the seriousness of participants was provided by the comparison of voting intentions for the German federal election in 2009 and self-reported voting behavior in the previous election in 2005. Participants who did not vote in 2005 or did not report their vote were excluded from this analysis. For the five parties detailed in Table 3, the agreement of the reported voting intention for the 2009 election and the vote cast in the previous election in 2005 was 66.7 % for serious ($n = 2{,}507$) and 51.5 % for nonserious ($n = 68$) participants. A z test for independent proportions showed this difference to be significant, $z = 2.61$, $p < .01$. Given the considerable stability of individual voting decisions (Schmitt et al., 2009), this result indicates that serious participants provided more valid answers than did nonserious participants.

As an additional test of data validity, the cumulative deviations of self-reported voting intentions from the official final election result were computed for all parties (Table 3). Note that predictions for the Conservative party underestimated its

Table 3 Cumulative deviations of serious and nonserious participants' self-reported voting intention for the 2009 German federal election from the final official result

	Official result	Serious		Nonserious	
	Percent	Percent	Deviation	Percent	Deviation
Conservatives	33.80	22.11	11.69	14.29	19.51
Social Democrats	23.00	23.33	0.33	29.46	6.46
Liberals	14.60	15.33	0.73	16.07	1.47
Left	11.90	16.99	5.09	11.61	0.29
Greens	10.70	12.88	2.18	12.50	1.80
Total	94.00	90.65	20.02	83.93	29.54

percentage of votes in the actual election by the largest margin (11.7 % for serious and 19.5 % for nonserious respondents), while the percentage of votes for all smaller parties—Liberals, Left, and Greens—was overestimated. An obvious explanation for this finding is that the higher age groups were underrepresented in our sample. It is the elderly population, however, from which most conservative voters are drawn (Federal Returning Officer, 2009). The difference between the cumulative deviations for serious (20.0 %, n = 3,378) and nonserious (29.5 %, n = 112) participants was found to be significant using a z test for two independent proportions, $z = 2.46$, $p < .02$. This finding seems to add further evidence for the notion that serious participants submitted more valid data than did nonserious participants.

However, before drawing a firm conclusion and following the suggestion of an anonymous reviewer, we tested a potential alternative explanation for this finding. Excluding nonserious respondents may spuriously improve the election forecast if, regarding their demographic makeup, nonserious respondents are less representative of the general population than are serious respondents. There were indeed some demographic differences between the two groups. First, there was a higher proportion of male respondents among the serious respondents (76.44 %) than among the nonserious respondents (64.29 %), $\chi^2(1, n = 3{,}490) = 8.13$, $p < .01$, and there was also a higher percentage of people having obtained a high school diploma (the German "Abitur") among serious (63.23 %) than among nonserious (43.75 %) respondents, $\chi^2(1, n = 3{,}490) = 16.75$, $p < .001$. We therefore used cell weighting (Kalton & Flores-Cervantes, 2003) to make the two groups comparable to the general population in terms of their sex and their level of education. The difference between the serious and the nonserious respondents regarding the cumulative deviation of their predictions from the final election outcome was thus somewhat reduced (from 9.52 % to 7.92 %), but it was still significantly larger than zero, $z = -1.93$, $p < .03$, one-tailed. Thus, the improved election prediction that resulted from the exclusion of the nonserious respondents could not be explained solely on the basis of a different demographic makeup of this group.

Nondemographic measures showed no significant differences between serious and nonserious respondents. Participants who failed the seriousness check took as much time to complete the survey (M = 574.19 s) as those who passed the check (M = 607.93 s), $t(204.16) = -1.29$, $p = .21$, and the rate of inconsistent answering was also comparable between the serious (5.09 %, n = 3,378) and the nonserious (5.36 %, n = 112) groups, $z = 0.85$, $p = .39$. Multiple participations were also as common among self-declared serious (1.59 %, n = 3,378) as among nonserious (1.79 %, n = 112) participants, $z = 1.53$, $p = .13$.

Benefits of competing screening methods

The utility of the seriousness check has to be compared with that of other methods of participant screening. Following the suggestion of an anonymous reviewer, we analyzed the beneficial effect of competing approaches on screening out participants. To this end, we computed the accordance of votes cast in 2005 and reported voting intentions for 2009, as well as the cumulative deviations between the voting percentages predicted for all parties on the basis of our survey and their actual results in the election. These two indices provided a measure of the validity gains that can be obtained by employing various screening procedures. First, we tested responses from unique IP addresses against multiple submissions from common IP addresses. The agreement of votes cast in 2005 and intended votes in 2009 was comparable between data sets with unique (66.3 %, n = 2,528) and common (66.0 %, n = 47) IP addresses, $z = 0.04$, $p = .97$. The cumulative deviation of the predicted election outcome from the official result was 20.4 % for unique IP addresses (n = 3,434) and 18.8 % for multiple submissions (n = 56). This difference was insignificant, $z = 0.29$, $p = .77$. Thus, multiple submissions originating from common IP addresses did not provide less valid data than did unique submissions.

To test the effect of hasty answering on data validity, we compared the fastest 10 % with the remaining 90 % of respondents. Again, the agreement of votes in 2005 and intended votes in 2009 was comparable between the fastest 10 % (66.8 %, n = 220) and the remaining (66.2 %, n = 2,355) participants, $z = 0.19$, $p = .85$. The prediction of the

official result of the election was only slightly less accurate when based on the fastest respondents (cumulative deviation: 23.9 %, n = 336) rather than on all other respondents (19.9 %, n = 3,154), z = 1.72, p < .09. Thus, screening out the fastest 10 % of participants had only a marginal effect on data validity.

We also compared participants providing consistent with those providing inconsistent responses. We considered answers inconsistent if participants declared a joint household income lower than their individual outcome. Whereas consistent respondents (n = 2,458) exhibited a 66.4 % accordance between their votes cast in 2005 and their voting intention in 2009, inconsistent respondents (n = 117) exhibited an accordance of 63.3 %. This difference was not significant, z = 0.70, p = .48. However, predicting the final official election outcome resulted in a significantly smaller cumulative deviation for consistently answering (19.6 %, n = 3,312) than for inconsistently answering (44.4 %, n = 178) participants, z = 7.92, p < .01. Respondents providing inconsistent answers to questions regarding their income thus provided less valid data regarding their voting behavior, too.

Incremental validity

The analyses above have shown that competing approaches are less successful in screening out invalid data than is a seriousness check. Even more important than the utility of employing a single screening criterion is its incremental validity, however. We therefore explored the incremental validity of the seriousness check—specifically, the usefulness of employing a seriousness check when applied after a unique IP check, a completion time check, or a consistency check had already been performed. We found that in every case, the seriousness check provided an added benefit to data quality. When we restricted the analyses to responses from unique IP addresses, the agreement of votes cast in 2005 and intended votes in 2009 was higher when participants reported being serious (66.7 %, n = 2,462) rather than nonserious (50.0 %, n = 66), z = 2.83, p < .01. The cumulative error in predicting the outcome of the election was also higher for serious (20.1 %, n = 3,324) than for nonserious (30.1 %, n = 110) respondents, z = 2.56, p < .02. Similarly, when applying the seriousness check after the exclusion of the fastest 10 % of participants, the agreement of the voting intention for the 2009 election and the vote cast in 2005 was higher for serious (66.7 %, n = 2,292) than for nonserious (47.6 %, n = 63) participants, z = 3.16, p < .01. The cumulative deviation from the official final election result, too, was smaller for serious (19.5 %, n = 3,050) than for nonserious (32.3 %, n = 104) respondents, z = 3.20, p < .01. After screening out inconsistent answers, the agreement of the reported voting intention in 2009 and the vote cast in 2005 was still higher for serious (66.9 %, n = 2,394) than for nonserious (48.4 %, n = 64), participants, z = 3.08, p < .01, and predicting the final official election outcome resulted in a smaller cumulative deviation for serious (19.3 %, n = 3,206) than for nonserious (30.4 %, n = 106) responses, z = 2.84, p < .01.

For a final comparison of serious and nonserious responses, we excluded participants on the basis of multiple IP addresses, fast completion times, and inconsistent answers. We were surprised to find that even after all competing measures had been applied in combination, the agreement of votes cast in 2005 and voting intention in 2009 was, again, higher when participants reported having answered seriously rather than nonseriously (66.9 %, n = 2,154 vs. 44.1 %, n = 59), z = 3.65, p < .01. Furthermore, the cumulative deviation for serious participants was 19.1 % (n = 2,860), as compared with 32.8 % for nonserious respondents (n = 98), z = 3.36, p < .01. Thus, we found evidence for the incremental utility of screening out nonserious participants; their exclusion improved data quality even when other frequently used screening criteria had already been applied.

Discussion

To test whether participants failing the seriousness check provide data of lesser quality than those passing the check, three indicators of data validity were examined. Serious participants' self-reported political attitudes on a left–right scale correlated significantly higher with their sympathies toward a 2009 government participation of the Social Democrats, the Liberals, and the right-wing coalition. For Conservatives, the Greens, and the left-wing coalition, this effect was also present but only marginally significant. The Left party was the only one for which no effect of seriousness was found. The accordance of voting intention for 2009 and voting behavior in 2005 was significantly higher for serious participants, and deviations from the official final election result were significantly smaller for respondents answering seriously.

It can therefore be concluded that as compared with serious participants, nonserious participants exhibited a significantly different and less reliable answering behavior. The exclusion of participants failing a seriousness check was thus shown to have the potential to considerably improve data validity. Moreover, we found strong evidence for the incremental validity of the seriousness check. Restricting the analyses to serious participants benefited data quality even after the data had already been screened for multiple IP addresses, fast completion times, or inconsistent answers. Even when all of these measures were used in combination, the seriousness check was able to improve the validity of the data. Therefore, we recommend adopting the seriousness check as a standard item in online studies. It would also be

desirable to see this technique implemented as a default in software for creating Internet-based studies, as is already the case in *WEXTOR* (Reips & Neuhaus, 2002).

The seriousness check we employed was conducted after the completion of the survey. This was done because we felt that asking the question after the completion of the survey was more likely to indicate the true nature of the participation, since only a response after the completion of the survey can reflect a potential change of mind during participation. However, Reips (2002, 2008, 2009) has argued that asking a seriousness question already in the early stage of a survey may serve the additional purpose of increasing motivation and reducing the subsequent dropout rate. A potential drawback of employing a seriousness check already at the beginning is that it may signal that nonserious responses are being expected, thereby increasing the rate of nonserious participants.

Since previous seriousness checks have sometimes been employed prior to, and sometimes subsequent to, participation in a study (see Table 1), potential effects of the timing of seriousness checks should be investigated more closely in future research. The self-reported seriousness at the end of a survey may be argued to be related to a respondent's level of motivation. Oppenheimer et al. (2009) found no association between self-reported motivation at the end of a study and the tendency to not read instructions and concluded that self-report measures do not seem to reliably capture a participant's motivation. The results of our study appear to contradict this finding, because we found the self-reported seriousness of participation to be predictive of the validity of the data. One possible explanation for this divergence of findings is that the seriousness check refers to actual answering behavior, whereas Oppenheimer et al. may have induced a focus on hedonic aspects by asking respondents about their motivation. It is easy to conceive of situations in which motivation is not the only determinant of serious responding, as, for example, when a tax form has to be submitted. It is also important to note that the very large sample in our study allowed us to detect motivational differences with high statistical power.

Any procedure used to screen out participants in an attempt to reduce noise in online research runs the risk of excluding valid data along with the less valid responses. As was noted by an anonymous reviewer, it may therefore be of interest to know why participants declare their participation as nonserious. In future studies, researchers may want to ask nonserious participants to elaborate on their answer to the seriousness check. For example, participants might be given an opportunity to indicate whether they felt not competent enough to answer the questions, were answering carelessly because they were interested only in the questions or in the formatting of the questionnaire, or were answering less thoughtfully because they were in a hurry. It might not be prudent to screen out a participant simply because he or she does not feel competent enough to answer all questions, and it might also be important to detail exactly what a researcher considers nonserious responding, to ensure that only invalid data are being excluded. The fact that data quality improved in a number of different ways in the present study, however, suggests that more valid data seem to remain after self-declared nonserious submissions have been discarded.

An anonymous reviewer of this study suggested that it may be possible and helpful to distinguish between different types of nonserious responding. For the present survey, all participants were recruited during an active search related to the election and were not incentivized. As a consequence, their motivation and interest in the survey topic was likely to be above average, and there was no apparent reason for them to be dishonest about the nonseriousness of their participation. Under these conditions, the seriousness check improved data quality and may have helped to detect and remove participants and fellow researchers taking part out of curiosity. Participants receiving a financial compensation for their answers, however, will probably be reluctant to admit nonserious responding, out of concern that their incentive will be withheld or that they will not be admitted to future surveys. This may apply, for example, to surveys conducted on Amazon's Mechanical Turk (Buhrmeister, Kwang, & Gosling, 2011). Participants motivated solely by a financial compensation have a strong incentive to increase their revenues by providing hasty or random answers or taking part multiple times. In such cases, or in surveys for which participation is mandatory, it may be advisable to introduce additional measures to detect different variants of nonserious responding. For example, an instructional manipulation check may help to detect inattentive participants, consistency checks and completion time checks may help to detect hasty or careless responding, and unique IP checks may help to detect multiple participations.

A drawback of the use of a multitude of screening methods is that they increase the "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011). Investigators may be tempted to try different combinations of methods and to report only the most favorable results. For example, a completion time cutoff may be chosen in view of the outcome of a subsequent statistical test, rather than on the basis of a solid rationale. Post hoc trial-and-error screening inflates the probability of a type I error, however, and may artificially produce evidence either supporting or contradicting any given hypothesis (Simmons et al., 2011). To ensure the validity of research results, it is therefore important that researchers decide a priori on their exclusion criteria and, subsequently, adhere to these decisions. Seriousness checks encourage the a priori consideration of exclusion criteria, because they have to be deliberately built into the survey. Using seriousness checks may thus have a positive effect on the researchers' degrees of freedom, provided that researchers

commit to dropping all nonserious responses and refrain from combining this check with other post hoc data filters in an arbitrary manner.

The magnitude of the effect of excluding nonserious participants obviously depends on their relative number. In the present study, the amount of nonserious participants was low, as compared even with the lowest rates previously reported in the literature (5 % – 6 % in Musch & Klauer, 2002). This may be due to the high relevance of the topic of this survey, which was conducted close to the election day. However, given their different answer behavior, a larger number of nonserious participants would have considerably altered the results of the survey. We therefore recommend routinely employing seriousness checks to improve data quality in an effortless and economic way.

References

Buchanan, T., Heffernan, T. M., Parrott, A. C., Ling, J., Rodgers, J., & Scholey, A. B. (2010). A short self-report measure of problems with executive function suitable for administration via the internet. *Behavior Research Methods, 42,* 709–714.

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah: Erlbaum.

Couper, M. P. (2000). Web-based surveys: A review of issues and approaches. *Public Opinion Quarterly, 64,* 464–494.

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments, & Computers, 31,* 572–577.

Egermann, H., Nagel, F., Altenmüller, E., & Kopiez, R. (2009). Continuous measurement of musically-induced emotion: A web experiment. *International Journal of Internet Science, 4,* 4–20.

Federal Returning Officer (2009). *Wahl zum 17. Deutschen Bundestag am 27. September 2009. Heft 4: Wahlbeteiligung und Stimmabgabe der Männer und Frauen nach Altersgruppen [Elections for the 17th German Bundestag on September 27, 2009. Volume 4: Electoral participation of and votes cast by men and women by age groups]*.

Fricker, R. D., & Schonlau, M. (2002). Advantages and disadvantages of internet research surveys: Evidence from the literature. *Field Methods, 14,* 347–367.

Ihme, J. M., Lemke, F., Lieder, K., Martin, F., Müller, J. C., & Schmidt, S. (2009). Comparison of ability tests administered online and in the laboratory. *Behavior Research Methods, 41,* 1183–1189.

Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers, 31,* 433–438.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics, 19,* 81–97.

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods, 41,* 1–12.

King, W. C., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80,* 643–651.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107,* 852–884.

Krantz, J. H., & Dalal, R. (2000). *Validity of Web-based psychological research* (pp. 35–60). San Diego, CA: Academic Press.

Lahl, O., Göritz, A. S., Pietrowsky, R., & Rosenberg, J. (2009). Using the World-Wide web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods, 41,* 13–19.

Li, M. F., & Olejnik, S. (1997). The power of rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21,* 215–231.

Madsen, H.S. (1987). *Utilizing Rasch analysis to detect cheating on language examinations*. Retrieved from ERIC database. (ED287284).

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly, 72,* 914–934.

Mangan, M., & Reips, U.-D. (2007). Sleep, sex, and the web: Surveying the difficult-to-reach clinical population suffering from sexsomnia. *Behavior Research Methods, 39,* 233–236.

Musch, J., & Klauer, K. C. (2002). *Psychological Experimenting on the World Wide Web: Investigating Content Effects in Syllogistic Reasoning* (pp. 181–212). Seattle, WA: Hogrefe & Huber.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872.

Pappi, F. (2009). Regierungsbildung im deutschen Fünf-Parteiensystem [Government formation in the German five party system]. *Politische Vierteljahresschrift, 50,* 187–202.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

Reips, U.-D. (2000). *The Web Experiment Method: Advantages, Disadvantages, and Solutions* (pp. 89–117). San Diego: Academic Press.

Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology, 49,* 243–256.

Reips, U.-D. (2008). *How Internet-mediated research changes science* (pp. 268–294). Cambridge, MA: Cambridge University Press.

Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. *Human Vision and Electronic Imaging XIV, Proceedings of SPIE, 7240,* 724008.

Reips, U.-D. (2011). *Privacy and the disclosure of information on the Internet: Issues and measurement* (pp. 71–104). Warsaw: UKSW Publishing House.

Reips, U.-D., & Buffardi, L. (2012). Studying migrants with the help of the internet: Methods from psychology. *Journal of Ethnic and Migration Studies, 38,* 1405–1424.

Reips, U.-D., & Neuhaus, C. (2002). WEXTOR: A web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods, 34,* 234–240.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15,* 217–226.

Schmidt, W. C. (1997). World-wide web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers, 29,* 274–279.

Schmitt, H., Sanz, A., & Braun, D. (2009). *Motive individuellen Wahlverhaltens in Nebenwahlen: Eine theoretische Rekonstruktion und empirische Überprüfung [Motives of individual voting behavior in Nebenwahlen: A theoretical reconstruction and empirical revision]* (pp. 585–605). Wiesbaden, Germany: Verlag für Sozialwissenschaften.

Schmitt, N., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17,* 143–150.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and

analysis allows presenting anything as significant. *Psychological Science, 20,* 1–8.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science, 280,* 867–873.

van den Wittenboer, G., Hox, J., & de Leeuw, E. (1997). *Aberrant Response Patterns in Elderly Respondents: Latent Class Analysis of Respondent Scalability* (pp. 155–162). Münster, Germany: Waxman.

Whitty, M. T., & Buchanan, T. (2010). What's in a screen name? attractiveness of different types of screen names used by online daters. *International Journal of Internet Science, 5,* 5–19.