Nuisance or Remedy? The Utility of Stylistic Responding as an Indicator of Data Fabrication in Surveys

Article *in* Methodology European Journal of Research Methods for the Behavioral and Social Sciences · January 2014

DOI: 10.1027/1614-2241/a000078

CITATIONS
15

READS
232

2 authors:

Christoph J Kemper
HSD University of Applied Sciences, Cologne
66 PUBLICATIONS **830** CITATIONS

SEE PROFILE

Natalja Menold
Technische Universität Dresden
71 PUBLICATIONS **499** CITATIONS

SEE PROFILE

Original Article

Nuisance or Remedy? The Utility of Stylistic Responding as an Indicator of Data Fabrication in Surveys

Christoph J. Kemper[1] and Natalja Menold[2]

[1]IMPP, Mainz, Germany, [2]GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

Abstract. Stylistic responding is usually seen as a nuisance by researchers working with questionnaire data due to its contaminating effects on the measurement of substantiative constructs. We demonstrate that stylistic responding may be useful to improve the data quality in surveys by allowing for an identification of deviant interviewer behavior – data fabrication – in survey fieldwork. Stylistic responding in $N = 710$ genuine and corresponding falsified interviews was compared. Genuine survey data was collected in paper-assisted personal interviews. Corresponding falsified data were obtained by instructing falsifiers to fabricate data based on person descriptions of genuine survey respondents. Acquiescent and midpoint responding, response range, and self-enhancement emerged as useful predictors of falsification. These indicators might now be used to develop and refine multivariate statistical methods for the ex-post identification of cheating interviewers in survey fieldwork.

Keywords: response sets, response bias, response style, interviewer falsification, cheating

Questionnaires using rating scales are an important method for collecting research data in the behavioral and social sciences. When responding to an item in a questionnaire, people do not only react to the substantiative meaning of the item but also to characteristics not related to item content. This non-content based form of responding is referred to as a response bias (Baumgartner & Steenkamp, 2001; Paulhus, 1991). By introducing an additional source of systematic variance, response biases may contaminate the measurement of substantiative constructs and threaten the validity of conclusions based on the data (for an overview see Ziegler, MacCann, & Roberts, 2011). A nuisance to most researchers working with questionnaire data, this systematic source of variance may be useful to reduce the impact of another source of contamination. Several survey researchers have put forward the idea that response biases can be used to improve data quality in surveys by identifying interviewers who are supposed to collect valid survey data in face-to-face interviews but cheat instead (Bredl, Winker, & Kötschau, 2008, 2012; Kemper, Trofimow, Rammstedt, & Menold, 2011; Menold, Winker, Storfinger, & Kemper, 2013; Schäfer, Schräpler, Müller, & Wagner, 2005). Large-scale surveys, for example, the General Social Survey (GSS), the Programme for International Student Assessment (PISA), or the German Socio-Economic Panel (SOEP), are an important source of research data in the social and behavioral sciences. The purpose of this article is to find indicators sensitive to falsification of survey data, thereby enabling researchers to develop and refine multivariate statistical procedures for the ex-post identification of cheating interviewers and to improve data quality in surveys.

According to Paulhus (1991), a response bias is a systematic tendency to respond to questionnaire items on some basis other than the specific content. This systematic tendency might be a temporary reaction to a situational demand (response set) or a stable and consistent behavior (response style). When responding to multi-point rating scale items measuring attitudes, behavior, or personality, some persons prefer extreme response categories of a rating scale (extreme responding) while others prefer moderate categories (midpoint responding) to describe themselves. Some persons tend to agree or disagree to items regardless of item content (acquiescent responding; disacquiescent responding). Persons may also differ in their variability when responding. While some persons use a wide range of the response categories around their mean response, others use only a narrow range (response range or non-differentiation). When responding to rating scale items, persons tend to give overly positive self-descriptions and may deliberately distort their responses in the face of situational demand (self-enhancement). For example, persons tend to emphasize positive and reject negative qualities of their personality (Paulhus, 2002) or claim knowledge about non-existent items (overclaiming; Paulhus, Harms, Bruce, & Lysy, 2003). Numerous studies have shown that these biases[1] occur when persons respond to rating scale items

[1] In the literature, some of those biases are classified as response sets while others are classified as response styles. To simplify matters, we do not distinguish these concepts and use the term stylistic responding to refer to both concepts.

DOI: 10.1027/1614-2241/a000078

(Baumgartner & Steenkamp, 2001; Kemper, 2010; Paulhus, 2002; Rammstedt & Kemper, 2011).

Stylistic responding is a major concern of researchers and practitioners using questionnaire data, because it introduces variance irrelevant to the primary purpose of the measurement – a systematic measurement error. The resulting contamination of the measures may bias statistical analyses thereby threatening the validity of conclusions based on the data. On the individual level, a respondent's score on a measure can be inflated or deflated by stylistic responding with detrimental effects on selection procedures, for example, in personal or treatment selection. On a group level, group means may be affected. Moreover, variances and covariances may be affected as well biasing multivariate analyses by increasing or decreasing correlations between measures. In the literature, abundant empirical support for the bias in mean and covariance structures due to stylistic responding can be found (e.g., Baumgartner & Steenkamp, 2001; Rammstedt, Kemper, & Borg, 2012; Ziegler, Toomela, & Bühner, 2009).

A crucial issue in research on stylistic responding is its measurement. There are basically two approaches[2] to measure stylistic responding. The first approach is to use scales specifically developed to measure certain forms of stylistic responding, for example, the L-Scale, F-Scale, or K-Scale from the Minnesota Multiphasic Personality Inventory-2 (Hathaway, McKinley, & Engel, 2000), the Crowne-Marlowe-Scale (Crowne & Marlowe, 1960), or the Balanced Inventory of Desirable Responding (Paulhus, 1998). However, this approach is often criticized, as these scales confound variance due to stylistic responding and variance due to a substantiative construct (Paulhus, 2002). Thus, some researchers propose to construct measures of stylistic responding which are unconfounded by substantiative content (Baumgartner & Steenkamp, 2001; Greenleaf, 1992). To achieve this goal, they use a set of many items stemming from scales measuring different constructs. Optimally, these items should be as heterogeneous in content as possible and, as a consequence, only marginally intercorrelated. The underlying assumption is that items that are only marginally intercorrelated minimize the contamination of stylistic responding measures with content-related variance. "... a respondent may agree or disagree with, respond extremely to, or endorse the midpoint on some items, but it is unlikely that across many items varying in content, the respondent's true position is characterized by consistent agreement or disagreement, extremity, or MPR [midpoint responding]" (Baumgartner & Steenkamp, 2001, p. 144). Using a set of only marginally intercorrelated items to construct stylistic responding measures is an approach widely applied.

Due to its effects on data quality, stylistic responding is usually seen as a nuisance to researchers working with questionnaire data. However, stylistic responding may also have a positive side. Several survey researchers propose that indicators of stylistic responding can be used to improve data quality in surveys (Bredl et al., 2008, 2012; Kemper et al., 2011; Menold et al., 2013; Schäfer et al., 2005). Data quality in surveys can be affected when interviewers who are supposed to collect valid survey data in face-to-face interviews intentionally depart from the designed guidelines or instructions and fabricate interview data (AAPOR, 2003). This deviant behavior of survey interviewers may occur with varying frequency (Biemer & Stokes, 1989; Bredl et al., 2008, 2012; Schräpler & Wagner, 2003) and may seriously affect statistical analyses, especially multivariate analyses and conclusions drawn from the data (Reuband, 1990; Schnell, 1991; Schräpler & Wagner, 2003). Thus, finding indicators sensitive to falsification is a critical research endeavor enabling the development of statistical methods for the ex-post identification of falsifiers and the improvement of data quality in surveys.

The basic idea linking stylistic responding to the improvement of data quality in surveys is (a) that real survey respondents (interviewees) differ from cheating interviewers (falsifiers) in motivational and cognitive processes when responding to a survey questionnaire, (b) that these interindividual differences manifest in observable differences in stylistic responding which could (c) be used to identify falsifiers of survey data (Kemper et al., 2011; Menold et al., 2013). Concerning motivational processes, we assume that falsifiers invest effort to select answers which minimize the chance of being detected whereas interviewees tend to give plausible rather than optimal answers (Krosnick, 1991). Concerning cognitive processes, we assume that falsifiers use implicit theories of personal behavior and stereotypes to infer most likely answers of putative interviewees (Borkenau & Liebler, 1992; Reuband, 1990; Schnell, 1991) whereas interviewees usually give honest answers when they respond to survey items. Moreover, automatic cognitive processes which are present when survey items are answered by interviewees (Sudman, Bradburn, & Schwarz, 1996; Tourangeau, Rips, & Rasinski, 2000) may be interrupted for falsifiers due to their high cognitive task engagement.

These interindividual differences in cognitive and motivational processes could be related to observable differences in stylistic responding. For example, we assume that falsifiers use common stereotypes to infer most likely answers of putative interviewees and respond to items more carefully resulting in less extreme and more moderate ratings. Consequently, an overall decrease of response range compared to the ratings of interviewees should occur. Moreover, we expect that falsifiers claim more familiarity of nonexistent items compared to interviewees, as falsifiers do not recognize the distracting nature of these items and falsify them based on the implied knowledge of the putative interviewee. We also assume that falsifiers' higher task engagement could result in less acquiescent responding on scales with a balanced set of items, as item content and the need for consistency become more salient. These are some assumptions on the mechanisms linking differences in cognitive and motivational processes to differences in stylistic responding. However, these underlying mecha-

[2] Within the framework of classical test theory.

nisms are not the focus of the research presented here. In this article, we rather focus on observed differences in stylistic responding found in previous studies and their utility for the discrimination of falsified from genuine interviews than on the underlying mechanisms causing these differences.

In previous research (Kemper et al., 2011), we demonstrated that differences in stylistic responding between interviewees and falsifiers do occur. We compared genuine interviews from the fieldwork of a large-scale survey and corresponding interviews artificially created – falsified – in our lab. As genuine interviews, a subset of 101 interviews from the German sample of the International Social Survey Program (ISSP) and the German General Social Survey (ALLBUS) of 2008 was used. Based on 11 sociodemographic characteristics of the survey respondents such as sex, age, education, and party preference, person descriptions were created. These person descriptions and the same questionnaire used in the ISSP/ALLBUS survey were given to the falsifiers in our lab. Falsifiers were instructed to fill out the questionnaire as the person described – the putative interviewee – would have done it. Indicators of extreme, midpoint, acquiescent, and disacquiescent responding were constructed from responses to the Big Five Inventory-10 (BFI-10; Rammstedt & John, 2007). For falsifiers, lower extreme and higher midpoint responding were observed in ratings compared to interviewees. Additionally, falsifiers demonstrated lower acquiescent and higher disacquiescent responding. In the same study, we could also demonstrate that these differences in stylistic responding may be used to identify falsifiers within the dataset. In multivariate analysis we were able to identify numerous falsifiers. Overall, 65.3% of interviews were correctly classified as either genuine or falsified. These results corroborate the utility of stylistic responding measures for the identification of falsifiers in survey datasets (cf. also Bredl et al., 2008, 2012; Schäfer et al., 2005). However, these results have not been replicated so far.

Aim of the present article is to replicate and extend previous findings (Kemper et al., 2011) concerning the utility of stylistic responding measures for the identification of deviant interviewer behavior. These previous findings were obtained comparing a subset of interviews taken from the German ISSP/ALLBUS survey of 2008 and corresponding falsified interviews. Falsifiers who fabricated interviews in our previous study differed in their experience of conducting genuine face-to-face interviews. As this kind of experience might play a role in the quality of falsifications, we decided to increase experimental control of this variable in the study presented here. Thus, we used a sample of students whom we allowed to gather experience by conducting genuine interviews in a first step before instructing them to fabricate interviews based on person descriptions in a second step. Moreover, by using a sample of students we were able to increase sample size considerably compared to our previous study, thereby increasing statistical power for multivariate analyses. Additionally, we introduced another crucial factor in survey fieldwork – interviewer payment (cf. Crespi, 1945).

To the best of our knowledge, the study presented here is the first carefully controlled experiment comparing genuine and fabricated interview data using properly trained, experienced, and motivated persons as falsifiers (for details see Method section). Moreover, to extend previous research we introduced additional indicators of stylistic responding. Besides extreme, midpoint, acquiescent, and disacquiescent responding used in the previous study, we also constructed a measure for response range or nondifferentiation, and used three measures of self-enhancement – exaggerating positive qualities, minimizing negative qualities, and overclaiming.[3] Due to conceptual reasoning and previous research reported above, we expected to find differences in stylistic responding between falsifiers and interviewees. We assumed lower extreme responding, higher midpoint responding, and an overall decreased response range for falsifiers. Moreover, we expected differences in self-enhancement scores between falsified and real interviews, as falsifiers have to infer the level of self-enhancement of the putative interviewees based on stereotypes and implicit theories of behavior. Concerning the second design factor – interviewer payment – we expected effects on motivational processes of falsifiers at least for some stylistic responding indicators. However, we had no specific hypotheses on main or interaction effects of payment as this issue is uncharted territory.

Method

Sample and Procedure

To obtain and compare genuine and falsified survey data, a two-step approach was followed. In a first step, we collected data from $N = 710$ survey respondents – the interviewees. The interviewees were students enrolled at the University of Giessen in 2012 (age Md = 24 years, IQR = 3; 60% females). These interviewees were recruited on the campus and interviewed by 78 interviewers who were students of the University of Giessen as well (age Md = 24 years, IQR = 3; 59.2% female). On average, each interviewer conducted 9.1 ($SD = 1$) interviews. Most interviews were conducted at home (58.1%) or on the university campus (28.4%). Each interview was recorded and checked after study completion to assure that interviews with survey respondents were actually conducted and not falsified. For conducting the interviews, interviewers received a payment either per interview (€8) or per hour (€12).

[3] Even though exaggerating positive qualities (EPQ) and overclaiming (OC) seem semantically overlapping, different measurement approaches apply (for details cf. Paulhus et al., 2003). The EPQ score is obtained by summing responses to items describing socially desirable behaviors, for example, being attentive all the time when talking to someone. The OC score is calculated from ratings of familiarity about existent and nonexistent items, for example, words and nonwords.

In a second step two weeks later, we fabricated survey data in the lab corresponding to data gathered in the field. Person descriptions based on the characteristics of the interviewees were composed. Following the approach of Reuband (1990), we selected several interviewee characteristics that a potential falsifier could easily obtain by a short interview, for example, sex, age, subject of study, number of semesters enrolled, place of residence, and country of origin. Afterwards, we invited the interviewers to our lab and instructed them to falsify survey data based on these person descriptions. They were briefed on the purpose of the study, that is, the improvement of methods to identify falsifications in survey data, their role as falsifiers, and the procedure. Each falsifier received the same materials he conducted the interviews with previously. We gave the person description of the putative interviewee to them and instructed them to imagine conducting a face-to-face interview with the person described. We instructed the falsifiers to fill out the survey questionnaire as the person described would have done it and to give answers that closely matched her answers. By this approach, we obtained $N = 710$ falsified interviews for $N = 710$ corresponding genuine interviews. In order to ensure that falsifiers did not fabricate a survey questionnaire of a person they had interviewed themselves, interviewee descriptions were randomly assigned to falsifiers and reassigned when necessary. Falsifiers were either paid per interview (€3) or per hour (€9). An incentive of 3×100 € (lottery) was offered for falsifiers who "cannot be detected by our statistical identification method." After fabricating interviews, falsifiers were instructed to fill out the survey questionnaire describing themselves.

Questionnaire

The survey questionnaire contained 62 questions on sociodemographics, attitude, behavior, and personality. Some of these measures are routinely used in the German General Social Survey (ALLBUS) and cover different domains, for example, attitudes on political issues, sex roles, the economic situation, and social justice. To accommodate the questionnaire to our target population, we added measures from a survey on the conditions of students in German universities (Isserstedt, Middendorff, Kandulla, Borchert, & Leszczensky, 2009). Finally, we used multi-item measures typically used for personality assessment – the BFI-10 (Rammstedt & John, 2007), the Vocabulary and Overclaiming Test (VOCT; Ziegler, Kemper, & Rammstedt, 2013), and items from a self-enhancement scale measuring the exaggeration of positive qualities and the minimization of negative qualities (Kemper, Beierlein, Bensch, Kovaleva, & Rammstedt, 2012).

Indicator Construction

Indicators were either constructed from responses to rating scale items of diverse multi-item multi-point attitude, behavior, and personality measures or from scales specifically developed to measure certain forms of stylistic responding. The former indicators of stylistic responding were constructed following common recommendations in the response style literature (cf. Baumgartner & Steenkamp, 2001; Kulas & Stachowski, 2009). For extreme responding (ERS), we counted the most extreme responses on the ratings scale, for example, "1" and "5," on a five-point rating scale, across all items and measures. For midpoint responding (MPR), we used only those measures with an uneven number of response categories and counted middle category endorsement. Acquiescent responding (ARS) and disacquiescent responding (DARS) were estimated based on BFI-10 responses. The BFI-10 is a balanced scale. It contains five pairs of items which share substantive content but differ in keying direction. ARS/DARS was constructed by counting the total number of agreement/disagreement responses regardless of keying direction. To measure (the restriction of) response range (RR), we calculated the standard deviation of respondent's responses across items. More specifically, standard deviations were calculated for sets of items with the same response scale, inverted, standardized, and averaged. To avoid missing values on these stylistic responding indicators, counts were divided by the number of items with valid responses. Concerning the indicators of self-enhancement, we calculated the scale scores of exaggerating positive qualities (EPQ), minimizing negative qualities (MNQ), and overclaiming (OC) according to the recommendations given by the scale developers (Kemper et al., 2012; Ziegler et al., 2013).

Statistical Analysis

Data were analyzed with SPSS 20. In order to test whether stylistic responding indicators are sensitive to falsification and differences in payment, we conducted a between-subjects Multivariate Analysis of Variance (MANOVA) with group membership (falsifiers vs. interviewees) and payment (payment per interview vs. per hour) as independent variables. MANOVA was followed by separate ANOVAs for the individual indicators. Control for familywise Type I error of ANOVAs was assured by prior MANOVA. Significance of effects was tested at $\alpha = .05$. Partial η^2 was used to indicate the size of effects. All effect sizes including correlation coefficients were evaluated according to the recommendations of Cohen (1988). Afterwards, we applied linear discriminant analysis (LDA) in order to examine the predictive power of the model. Stylistic responding indicators as well as payment condition and significant interaction terms, ERS × payment, RR × payment, and MNQ × payment, were used to predict group membership. To avoid biased estimates, we applied a cross-validation approach with jackknife classification (cf. Tabachnick & Fidell, 2007). Overall hit ratio, sensitivity, and specificity were calculated.

Results

To investigate the effect of falsification and payment condition on the combined dependant variables, SPSS

MANOVA was conducted. With the use of the Wilk's λ criterion, a significant effect of falsification was observed, $F_{(8, 1409)} = 28.00$, $p < .001$, $\eta_p^2 = .137$. According to η_p^2, there was a strong association between the combined indicators and group membership (cf. Cohen, 1988). Thus, falsifiers differed substantially from interviewees with respect to the combination of stylistic responding indicators used. For payment, MANOVA yielded a small but substantial effect, $F_{(8, 1409)} = 3.37$, $p < .01$, $\eta_p^2 = .019$, suggesting that response behavior of interviewees and falsifiers differs depending on the remuneration scheme. The interaction term was significant as well, $F_{(8, 1409)} = 2.52$, $p < .01$, $\eta_p^2 = .014$.

To investigate the effect of falsification and payment on the individual dependant variables, separate ANOVAs were conducted. Results are depicted in Table 1. Substantial interaction effects of falsification and payment were found for ERS, RR, and MNQ. The type of interaction implied that an unequivocal interpretation of the main effects of falsification is possible for these indicators (cf. Bühner & Ziegler, 2009). However, due to a hybrid interaction an interpretation of the main effect of payment is not possible for RR. Concerning main effects of falsification, substantial differences between interviewees and falsifiers were found for all stylistic responding indicators except DARS. In falsified compared to genuine interviews, less ERS and more MPR occurred. When falsifying the study questionnaire, falsifiers apparently choose less extreme categories of rating scales and more midpoint categories. This pattern of results is consistent with a lower response range (RR) of the falsifiers compared to interviewees. RR yielded the strongest effect of falsification. ARS yielded an effect of similar strength with falsifiers showing less ARS compared to interviewees. Finally, we observed substantial differences in three measures of self-enhancement, EPQ, MNQ, and OC, with falsifiers attributing a stronger tendency for an overly positive self-description to putative interviewees. Concerning the impact of payment, the significant main effect for MNQ suggests a stronger rejection of negative qualities in interviews conducted on a per hour versus per interview remuneration scheme.[4]

To examine the predictive power of the model, LDA was applied. Predictors were stylistic responding indicators, payment condition, and significant interaction terms. Box's M indicated a violation of that assumption of equality of covariance matrices. However, given the large and equal group sizes, this problem is not regarded as serious. One discriminant function was calculated revealing a substantial association between predictors and group membership, $\chi^2 = 220.6$, $df = 12$, $p < .001$, accounting for 14% of between group variability (canonical $R^2 = .14$). The structure matrix of correlations between predictors and discriminant function suggests that the best predictors to distinguish falsifiers from interviewees are RR (.53), ARS (−.52), EPQ (.35), OC (.24), MPR (.20). Remaining stylistic responding indicators, payment condition, and interaction terms had negligible discriminant loadings. The cross-validated classification showed that overall 67.7% were correctly classified. Sixty-eight percent of falsifiers were correctly classified as falsifiers by the model (sensitivity) and 67.5% of interviewees (specificity). To examine the practical relevance of the payment condition and the interaction terms, we repeated the LDA without these predictors. Neither overall model fit (canonical $R^2 = .14$) nor hit ratio (66.7%) changed substantially. Thus, their discriminative power is marginal.

Discussion

Falsification of survey data by cheating interviewers is considered a problem for survey research, as it may contaminate data and threaten the validity of conclusions based on that data. Finding indicators which are sensitive to falsification is a necessary first step for the development of multivariate statistical procedures for the ex-post identification of cheating interviewers. In our research, we demonstrate that indicators of stylistic responding might be useful in this regard. We set out with the assumption that real survey respondents and interviewers who fabricate data differ in their cognitive and motivational processes when responding to survey questionnaires and proposed that indicators of stylistic responding might be used to measure these differences. In our study, expected differences in stylistic responding between interviewees and falsifiers were indeed observed.

Substantial differences in stylistic responding in ratings of attitude, behavior, and personality were found for all indicators proposed except disacquiescent responding. Falsifiers used less extreme categories, more midpoint categories and showed less acquiescent responding compared to interviewees. In a previous study (Kemper et al., 2011), the same pattern of results was observed. Moreover, acquiescent responding yielded the strongest group differences of the stylistic responding indicators in our previous study. In the current study, effects for this indicator were second in strength after response range. However, response range was not used in the previous study. Despite considerable differences between studies in sample composition and indicator construction results could be replicated. Our results are also in line with those of others concerning extreme responding. Bredl et al. (2008, 2012) found differences between falsifiers and interviewees for extreme responding as well. Thus, evidence accumulates supporting the utility of extreme, midpoint, and acquiescent responding to indicate group differences in response behavior.

Besides a mere replication of results observed previously, we extended our research by testing new indicators and (design) factors. Differences in stylistic responding

[4] Exploratory analyses into possible mechanisms of payment effects on interviewer behavior were also conducted. Interestingly, experimental conditions were only marginally different concerning the duration of interviews/falsifications – interviewer paid per hour (Md = 33, IQR = 9) versus per interview (Md = 31, IQR = 9), falsifier paid per hour (Md = 33, IQR = 9.5) versus per interview (Md = 32.5, IQR = 8).

Table 1. ANOVA results – effect of Falsification, Payment, and Falsification × Payment on stylistic responding indicators

	Mean (SE)				Effect of					
	Per interview		Per hour		Falsification		Payment		Falsification × Payment	
Indicator	Interviewee	Falsifier	Interviewee	Falsifier	$F_{(1,\ 1416)}$	η_p^2	$F_{(1,\ 1416)}$	η_p^2	$F_{(1,\ 1416)}$	η_p^2
ERS	.23 (.006)	.23 (.006)	.23 (.007)	.20 (.007)	6.22*	0.004	3.25	0.002	5.05*	0.004
MPR	.22 (.006)	.23 (.006)	.22 (.006)	.24 (.006)	9.45**	0.007	1.25	0.001	0.03	0
ARS	.52 (.007)	.47 (.007)	.51 (.007)	.46 (.007)	65.35***	0.044	2.38	0.002	0.64	0
DARS	.31 (.006)	.32 (.007)	.32 (.007)	.33 (.007)	3.06	0.002	1.05	0.001	0.25	0
RR[a]	−.10 (.026)	.04 (.026)	−.12 (.028)	.18 (.027)	69.92***	0.047	5.50*	0.004	8.85**	0.006
EPQ	3.41 (.030)	3.61 (.030)	3.47 (.031)	3.60 (.031)	29.07***	0.02	0.53	0	1.07	0.001
MNQ	3.48 (.038)	3.50 (.039)	3.54 (.041)	3.72 (.040)	6.57**	0.005	12.83***	0.01	3.99*	0.003
OC	.56 (.011)	.603 (.011)	.56 (.012)	.603 (.012)	13.57***	0.01	0.05	0	0.06	0

Notes. ERS = extreme responding; MPR = midpoint responding; ARS = acquiescent responding; DARS = disacquiescent responding; RR = response range/nondifferentiation; EPQ = exaggerating positive qualities; MNQ = minimizing negative qualities; OC = overclaiming; SE = standard error. [a]averaged standardized. $*p < .05$. $**p < .01$, $***p < .001$.

were also found for response range. The use of response range stems from the idea that falsifiers have a lower variability in ratings of attitude, behavior, and personality compared to interviewees (Schäfer et al., 2005). We were able to corroborate this assumption. In our study, response range of falsifiers was substantially lower compared to interviewees. Notably, the strongest group difference overall was observed for this indicator. Falsifiers chose less extreme and more moderate ratings resulting in a lower variability of ratings. For the time being, response range seems the most promising of all stylistic responding indicators pending replication.

Besides group differences in extreme, midpoint, and acquiescent responding, as well as response range, falsifiers and interviewees differed also with respect to (predicted) self-enhancement. Three measures of self-enhancement were applied in our study as indicators of stylistic responding. For falsifiers, higher scores of EPQ, MNQ, and OC were observed compared to interviewees. These results suggest that falsifiers generally overestimate the putative interviewees' tendency to self-enhance by exaggerating positive and minimizing negative attributes or by overclaiming knowledge. This result is consistent with findings reported by Krueger (1998). Krueger demonstrated that persons expect others to self-enhance but may not be accurate judges concerning the actual level of other's self-enhancement.

Significant effects on stylistic responding indicators were not only observed for falsification but also for the second design factor – payment – and the interaction of both factors. For example, negative qualities were rejected more strongly when payment was per hour and not per interview. Further, whereas interviewees extreme responding was largely unaffected by payment, extreme responding of falsifiers was lower when paid per hour. A corresponding result was observed for response range. Responses of falsifiers were more homogenous when paid by the hour as compared to payment by interview whereas responses of interviewees were not affected by the remuneration scheme of the interviewer. The payment and falsification × payment effects are interesting in their own right as well as the possible mechanisms underlying these effects, for example, duration of interviews/falsifications. They yield a good starting point to further explore the utility of monetary incentives impacting interviewer's behavior and performance. However, according to the LDA results reported here incentives may have only marginal practical significance. They did not improve the identification of falsifiers in our study above the discriminative power of stylistic responding indicators.

The work presented here represents an attempt to clarify the utility of stylistic responding indicators for the discrimination of genuine and falsified survey data. We demonstrated that these indicators can reliably identify falsified interviews with a hit ratio substantially better than what could be achieved by chance. However, some limitations have to be kept in mind. For example, our study was conducted in the lab under optimal conditions. Groups of falsifiers and interviewees were equal in size fostering the detection of group differences. In the field, the group of

falsifiers can be expected to be substantially smaller in most surveys (Biemer & Stokes, 1989; Schräpler & Wagner, 2003) probably attenuating group differences in stylistic responding. Another issue concerns the generalizability of stylistic responding indicators. Each survey questionnaire consists of a unique set of items. Whether exploiting the secondary source of systematic variation in rating scale items – the non-content based variation in stylistic responding – works equally well for different survey questionnaires cannot be taken for granted and has to be clarified in future studies. For extreme, midpoint, and acquiescent responding a first replication by our research group can be deemed successful.

In the research presented here, we demonstrated that stylistic responding indicators may help to differentiate falsified from genuine interviews. These indicators – in conjunction with other formal and content-related indicators (e.g., item nonresponse, filter or open-ended questions; Menold & Kemper, 2013; Menold et al., 2013) – might now be used to develop and refine multivariate statistical methods for the ex-post identification of cheating interviewers in surveys and to improve data quality by excluding fabricated data from survey datasets (cf. Bredl et al., 2008, 2012; Menold et al., 2013). The applicability of a highly efficient method for the identification of bad apples in the interviewer staff seems a long way down the road. However, a first step was made. As the assurance of the integrity of research data used in the behavioral and social sciences by preventing or controlling for interviewer falsification is a reasonable goal to pursue (AAPOR, 2003), further efforts are strongly encouraged.

Acknowledgments

We would like to thank the DFG for its financial support in Projects WI 2024/2-1 and ME 3538/1-1 within SPP 1292. We would also like to thank Peter Winker, Nina Storfinger, and Viktoria Trofimow for their helpful support in the research presented in this paper.

References

American Association for Public Opinion Research. (2003). *Interviewer falsification in survey research: Current best methods for prevention, detection and repair of its effects* Retrieved from http://www.aapor.org/Content/aapor/Resources/ForResearchers/falsification.pdf

Baumgartner, H., & Steenkamp, J. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research, 38*, 143–156.

Biemer, P. P., & Stokes, S. L. (1989). The optimal design of quality control samples to detect interviewer cheating. *Journal of Official Statistics, 5*, 23–39.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62*, 645–657. doi: 10.1037/0022-3514.62.4.645

Bredl, S., Winker, P., & Kötschau, K. (2008). *A statistical approach to detect cheating interviewers* (ZEU Discussion Paper, 39). Retrieved from http://econstor.eu/bitstream/10419/39808/1/593464877.pdf

Bredl, S., Winker, P., & Kötschau, K. (2012). A statistical approach to detect interviewer falsification of survey data. *Survey Methodology, 38*, 1–10.

Bühner, M., & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler* [Statistics for psychologists and social scientists]. Munich, Germany: Pearson.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Crespi, L. P. (1945). The cheater problem in polling. *Public Opinion Quarterly, 9*, 431–445.

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

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly, 56*, 328–351. doi: 10.1086/269326

Hathaway, S. R., McKinley, J. C., & Engel, R. (2000). *MMPI-2. Minnesota Multiphasic Personality Inventory 2*. Bern, Switzerland: Huber.

Isserstedt, W., Middendorff, E., Kandulla, M., Borchert, L., & Leszczensky, M. (2009). *Die wirtschaftliche und soziale Lage der Studierenden in der Bundesrepublik Deutschland 2009* [The economic and social situation of students in Germany 2009]. Hannover, Germany: BWH Retrieved from http://www.sozialerhebung.de

Kemper, C. J. (2010). *Das Persönlichkeitsmerkmal Angstsensitivität: Taxon oder Dimension?: Eine Analyse mit dem Mischverteilungs-Raschmodell* [Trait Anxiety Sensitivity: Taxon or Dimension? A Mixed Rasch Model Analysis]. Hamburg, Germany: Dr. Kovac.

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2012). *Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: Die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G)* [A short scale for the assessment of two aspects of socially desirable responding]. Cologne, Germany: GESIS (GESIS Working Papers 2012|25).

Kemper, C. J., Trofimow, V., Rammstedt, B., & Menold, N. (2011, August–September) Poster session presented at the 11th European Conference on Psychological Assessment, Riga, Latvia. Retrieved from http://www.christoph-kemper.net/fileadmin/user_upload/Poster/ Kemper_et_al.__2011__Faking_indicators_ BFI10_Survey_research_ECPA.pdf

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measurement in surveys. *Applied Cognitive Psychology, 5*, 213–236. doi: 10.1002/acp. 2350050305

Krueger, J. (1998). Enhancement bias in descriptions of self and others. *Personality and Social Psychology Bulletin, 24*, 505–516. doi: 10.1177/0146167298245006

Kulas, J. T., & Stachowski, A. A. (2009). Construct and response bias correlates in summated scale definitions of personality traits. In L. B. Palcroft & M. V. Lopez (Eds.), *Personality Assessment. New research* (pp. 259–274). New York, NY: Nova Science.

Menold, N., & Kemper, C. J. How do real and falsified data differ? Psychology of survey response as a source of falsification indicators in face-to-face surveys. *International Journal of Public Opinion Research*. Advance online publication. doi: 10.1093/ijpor/edt017

Menold, N., Winker, P., Storfinger, N., & Kemper, C. J. (2013). A method for ex-post identification of falsifications in survey data. In N. Menold, P. Winker, & R. Porst (Eds.), *Survey Standardization and Interviewers' Deviations – Impact, Reasons, Detection and Prevention* (pp. 25–48). Frankfurt, Germany: Peter Lang.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson (Ed.), *Measures of social psycho-*

logical attitudes: Vol. 1. Measures of personality and social psychological attitudes (pp. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L. (1998). *Manual for the Paulhus Deception Scales: BIDR Version 7*. Toronto, Canada: Multi-Health Systems.

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

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology, 84*, 890–904. doi: 10.1037/0022-3514.84.4.890

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*, 203–212. doi: 10.1016/j.jrp. 2006.02.001

Rammstedt, B., & Kemper, C. J. (2011). Measurement equivalence of the Big Five: Shedding further light on potential causes of the educational bias. *Journal of Research in Personality, 45*, 121–125. doi: 10.1016/j.jrp. 2010.11.006

Rammstedt, B., Kemper, C. J., & Borg, I. (2012). Correcting Big Five measurements for acquiescence: An 18-country cross-cultural study with representative samples. *European Journal of Personality, 27*, 71–81. doi: 10.1002/per.1894

Reuband, K. H. (1990). Interviews, die keine sind. ''Erfolge'' und ''Mißerfolge'' beim Fälschen von Interviews [Interviews which are no genuine interviews. „Success'' and „failiure'' in interview falsification]. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42*, 706–707.

Schäfer, C., Schräpler, J. P., Müller, H., & Wagner, G. G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Journal of Applied Social Science Studies, 125*, 183–193.

Schnell, R. (1991). Der Einfluss gefälschter Interviews auf Survey-Ergebnisse [The impact of falsified interviews on survey results]. *Zeitschrift für Soziologie, 20*, 25–35.

Schräpler, J. P., & Wagner, G. (2003). Identification, characteristics and impact of faked interviews in surveys: An analysis by means of genuine fakes in the raw data of SOEP,. (IZA Discussion paper series, No. 969). Retrieved from http://hdl.handle.net/10419/20205

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston, MA: Pearson and Allyn & Bacon.

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

Ziegler, M., Kemper, C. J., & Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T). *Journal of Individual Differences, 34*, 32–40. doi: 10.1027/1614-0001/a000093

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). *New perspectives on faking in personality assessment*. New York, NY: Oxford University Press.

Ziegler, M., Toomela, A., & Bühner, M. (2009). A reanalysis of Toomela (2003): Spurious measurement error as cause for common variance between personality factors. *Psychology Science Quarterly, 51*, 65–75.

Received June 15, 2011
Accepted September 3, 2013
Published online January 21, 2014

Christoph J. Kemper

Institute for Medical and Pharmaceutical
Proficiency Assessment
Große Langgasse 8
55116 Mainz
Germany
E-mail contact@christoph-kemper.net
Web: www.christoph-kemper.net

