mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology
Volume 15, 2021 | 2

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246526
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession‘s knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.
All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Layout: Bettina Zacharias (GESIS)
Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

methods, data, analyses | 2021, Vol. 15(2), pp. 125-160 DOI: 10.12758/mda.2021.02

Do Falsifiers Leave Traces? Finding Recognizable Response Patterns in Interviewer Falsifications

Sandra Walzenbach
University of Konstanz

Abstract

Fraud by interviewers is a ubiquitous threat to data quality in survey practice, whenever face-to-face surveys are conducted. Particularly if interviewers use stereotypes about respondents to fill in questionnaires, falsifications can limit the variety of possible answers, lead erroneously to significant correlations and distort survey results.

In addition to external control mechanisms to detect fraud (such as postcards or time stamps) more recent research has started to also consider internal indicators (such as the number of missing values or open answers) as a monitoring strategy. This latter approach relies on ex-post statistical analyses and implicitly assumes that falsifiers apply rational behavioral strategies which result in detectable response patterns. This study examines to what extent fieldwork monitoring can benefit from such approaches, by empirically assessing how effective different indicators are at detecting known cases of fabrication.

In contrast to most previous research, which often relies on laboratory fabrications, this study uses authentic cases of detected interviewer fraud from a survey on the fairness of earnings conducted in Germany. The main goal of this study is to examine to what extent the falsifiers' attempts to produce unsuspicious data led to recognizable response patterns. For this purpose, we test a wide range of indicators that could potentially identify falsifications: avoidance of extreme categories and open text-based answers, low rates of item-nonresponse, strategic use of filter questions to shorten the questionnaire and non-compliance of responses to numeric questions with Benford's Law. Furthermore, we compare authentic and fabricated interviews according to their values on a social desirability scale and report results from an innovative trick question that was especially designed to detect falsifiers.

Keywords: interviewer falsification, interviewer fraud, interviewer effects, response patterns, statistical methods, data quality

Interviewer Falsifications as an Omnipresent but Seldomly Discussed Topic

Whenever interviewers are employed to collect survey data, researchers have to face the problem that their interviewers do not necessarily follow the same interests as they do. In fact, the interviewers' goals might even contradict the researchers' in many aspects (Winker 2016). While researchers are interested in unbiased data, ideally obtained from random samples with high response rates, interviewers might, in the worst case, aim to waste as little time as possible to conduct the necessary amount of interviews. Delivering quality data might not be their primary concern. In his book on 19 years of professional experience as an interviewer, Dorroch reports how interviewers complete their daily tasks with the least investment of effort and time possible (Dorroch 1994). Apart from being a nightmare for every researcher, his narration clearly points out that, from an interviewer's perspective, deviating from the interviewer guidelines and falsifying data can be a very beneficial rational decision.

Estimates of the actual amount of fabricated data in surveys vary to some extent, but most authors assume a share of between less than one and seven percent (Finn & Ranchhod 2017; Koch 1995; Schräpler & Wagner 2003; Schnell 1991; Schreiner, Pennie, & Newbrough 1988), while some mention possible numbers of above 50 percent depending on the survey and its supervision capacities (De Haas & Winker 2016).

Although data falsification is an omnipresent problem in survey research and sometimes even receives some attention in popular media (as in a prominent feature by the German Spiegel magazine in February 2018), little scientific literature has focused on the topic. This is particularly surprising because previous research has shown that fabricated data can systematically bias survey results: Schnell (1991) quantifies the potential threat of falsifications for data quality by varying the share

Acknowledgements

A first version of this paper was written in Katrin Auspurg's seminar on "Survey Research Methods" at the University of Konstanz. I would like to express my gratitude for her support and guidance. The survey on the fairness of earnings was part of the project "The factorial survey as a method for measuring attitudes in population surveys", which received funding from the German Research Foundation (DFG). The survey was designed, organised and monitored by Carsten Sauer, Katrin Auspurg, Thomas Hinz, and Stefan Liebig. Thomas Hinz had the crucial idea for the trick question that is discussed on page 13.

Direct correspondence to

Sandra Walzenbach, University of Konstanz, Department of Sociology,
Universitätsstraße 10, 78464 Konstanz, Germany
E-mail: sandra.walzenbach@uni-konstanz.de

of (laboratory) fabrications in a data set. He concludes that a share of five percent hardly affects univariate analyses. Multivariate analyses, however, were much more susceptible to bias (also see Reuband 1990). This paper therefore aims to contribute to a debate that we consider necessary in order to find an adequate way of dealing with falsification by interviewers.

Survey agencies and researchers usually apply a variety of monitoring strategies to deal with potential interviewer fraud (AAPOR 2003; Murphy, Biemer, Stringer, Thissen, Day, & Hsieh 2016). For our purposes, it is sufficient to distinguish between what we will call external and internal control mechanisms:

- *External control mechanisms* are external to the substantive answers in the questionnaire. Widely used techniques include recontacting respondents via postcard or phone call (e.g. Koch 1995: 91f), the storage of paradata such as time stamps to determine the length of an interview (e.g. Hood & Bushery 1997: 820f) or the number of conducted interviews per day (Bushery, Reichert, Albright, & Rossiter 1999: 317f) and the time gap between them. Some authors suggest that more experienced interviewers might use more sophisticated forms of falsifications (Schreiner et al. 1988), a consideration that leads Hood & Bushery (1997) to keep track of suspiciously high numbers of ineligible households that the interviewer might have misclassified to avoid hard-to-reach respondents. Less common but promising observational methods such as audio recordings or GPS tracking also belong to this category (Thissen 2014; Thissen & Myers 2016; Wagner, Olson, & Edgar 2017).
- *Internal control mechanisms* refer to statistical ex-post analyses of the substantive answers from the questionnaire. In contrast to external control mechanisms, the analysis of internal response patterns aims to develop a technique that can identify falsifications merely on the basis of the completed questionnaires themselves (e.g. Bredl, Winker, & Kötschau 2012; De Haas & Winker 2016; Kosyakova, Olbrich, Sakshaug, & Schwanhäuser 2019). This more controversial approach draws on rational choice theory and the assumption that the falsifier's attempt to produce unobtrusive and unsuspicious data results in certain recognizable response patterns.

Internal control mechanisms are not meant to replace external checks. Rather, they are a cost-efficient supplement to external control mechanisms that can be useful to preselect suspicious interviewers for further, more targeted examination.

Research Objective and Approach: Do Falsifiers Leave Traces?

So far there is no scientific consensus on a superior monitoring strategy to deal with interviewer fraud (Murphy et al. 2016). Instead there is a variety of coexisting measures that either aim to prevent or retrospectively discover falsifications, particularly when it comes to internal control mechanisms, that is, *indicators* that are internal to the collected questionnaire data. Some authors have suggested principal component analysis to examine similar response patterns on ordinal response scales (Blasius & Thiessen 2013) or cluster analyses that combine several statistical indicators to identify interviewers at risk (e.g. Bredl et al. 2012). However, these latter approaches often suffer from a high number of false positives, particularly in settings where the individual interviewers complete few interviews and the share of fraudulent interviewers is low (De Haas & Winker 2014; Storfinger & Winker 2013) and no a priori restriction on the number of falsifiers is defined (De Haas & Winker 2016). In addition, it is unclear which indicators are best suited for cluster analyses (see Menold, Winker, Storfinger, & Kemper 2013 for a simulation study testing different combinations of indicators on a laboratory sample with 50% falsifications).

As will be argued in more detail in the following section, the little research that empirically tests such indicators has produced somewhat contradictory results. This paper therefore focuses on statistical ex-post analyses of response patterns and examines to what extent the response patterns in fabricated data reflect the typically assumed 'rational' interviewer behavior that translates into detectable peculiarities: Do falsifiers leave traces that make them identifiable?

We will empirically test five internal indicators using a survey on the fairness of earnings in Germany. These data are particularly suitable for the present research, because they contain (at least) 44 authentic cases of interviewer fraud, which can be tested for typical response patterns. Apart from more conventional external control mechanisms, such as control postcards and time stamps (by which these falsifications were discovered), the questionnaire also contained a trick question on the income inequality in Europe. This rather innovative attempt to identify falsifiers merely by their response patterns will be discussed in more detail later on.

A major limitation of most previous studies on interviewer falsifications is that they rely on artificial laboratory experiments, in which arbitrarily chosen respondents (often university students) are asked to fabricate data (for a recent exception see Schwanhäuser, Sakshaug, Kosyakova, & Kreuter 2020). It is a crucial advantage that the data at hand allow us to analyze data from authentic interviewers with intentions to fake data in a real life situation.

Previous Research

Types of Interviewer Fraud

Generally, interviewer fraud is defined as an intentional deviation from the interviewer guidelines (AAPOR 2003; Gwartney 2013). These deviations can occur at different steps of the interview process and vary in their degree of severity (for an extensive list see Murphy et al. 2016). The AAPOR (2003) talks about "a continuum of severity of falsification" (page 2). For the purpose of this study, milder forms of interviewer deviations such as rephrasing questions, failing to record verbatims or allowing refusal and item-nonresponse will not be discussed in more detail. Although focus group interviews among interviewers suggest that such minor deviations are the most common type of interviewer falsification (Nelson & Kiecker 1996), it can be hard to determine in an individual case if e.g. by rephrasing, an interviewer intended to help a respondent or to falsify data.

Leaving minor interviewer deviations aside, Schnell (1991) distinguishes between three essential types of falsifications, into which most other classifications (e.g. AAPOR 2003; Schreiner et al. 1988) can be condensed:

- *complete falsifications*, meaning that the interviewer fills in the whole questionnaire without contacting the designated respondent
- *partial falsifications*, for which the interviewer collects some crucial information, either directly from the respondent or from someone who knows him/her, in order to complete the remaining questions alone
- cases in which the *random procedure to select respondents is ignored*, meaning that instead of the target subject someone else is interviewed or the eligibility of a potential hard-to-reach respondent is misreported

The three types of interviewer fraud differ in two aspects: In how demanding it is (for the interviewer) to produce them and in how demanding it is (for the researcher) to detect them (Schnell 1991: 27-29). This last distinction is crucial for this paper insofar as falsifications that ignore the random selection process are impossible to discover solely by means of statistical ex-post analyses of the data. Real respondents will produce unsuspicious response patterns no matter if they were part of the random sample or not. Hints for this specific kind of fraud can only be gained with the aid of external control mechanisms. As a consequence, the approach presented here will only be helpful for the identification of certain types of falsifications. At the same time, the analyses of response patterns cannot be - and do not intend to be - a replacement for external control mechanisms but a first step to identify suspicious cases. Moreover, partial falsifications will be harder to identify than complete falsifications. For partial falsifications, detection will be easier to accomplish the larger the falsified fraction of questions within an interview (De Haas & Winker 2014).

Interviewer Fraud as a Rational Behavior

Fraud can be considered the result of a rational decision process, in which the interviewers react to the situational circumstances they encounter. According to subjective expected utility theory (Kroneberg & Kalter 2012; Esser 1999: 247-275), a wide version of rational choice theory, such a decision process is a function of the following factors:

- the alternative actions to choose from
- the subjective utilities associated with these alternatives
- the costs associated with each alternative
- the perceived probability that an action can actually be carried out and thus leads to the expected utility

The worst case scenario from a researcher's perspective would be an interviewer who aims to complete the job in as little time and with as little effort as possible. In line with theoretical assumptions, the respective survey literature has identified a variety of circumstances that might make fraud more likely. Gwartney (2013) argues that "calculating cynics" (page 203) who fake frequently and systematically are rare in practice. She believes that most interviewers rather fake occasionally, when their ethics break down in difficult situations. However, she acknowledges that the interviewers' working environment can strongly encourage them to deviate from instructions. Already decades ago, Crespi (1945) argued that researchers and survey agencies can change the "demoralising" circumstances, in which interviewers make their decisions. Similarly, Koch pointed out in the 1990s that it would be wrong to solely blame the interviewers. He considered defective interviewer training, long and poorly designed questionnaires, and meagre salaries a part of the problem (Koch 1995: 102). Gwartney (2013) adds complicated sampling procedures and software, performance and deadline pressures, and a lack of appreciation and support from fieldwork agencies to the list. Interesting empirical evidence for these problems is provided by a qualitative study among interviewers (Nelson & Kiecker 1996) and a field experiment that manipulates the interviewers' working conditions (Menold, Landrock, Winker, Pellner, & Kemper 2018). In addition, some recent work highlights the importance of work ethics and moral values which should be articulated by researchers and supervisory staff (AAPOR 2003; Gwartney 2013; Murphy et al. 2016).

Empirical Evidence on Suspicious Response Patterns

It has been suggested that interviewers do not only act rationally when they make the decision to (or not to) falsify, but also while they are trying to produce unsuspicious data: "Interviewers who falsify will try to keep it simple and fabricate a

minimum of falsified data" (Hood & Bushery 1997: 820). If this is the case, faked data would show statistically detectable differences to properly completed questionnaires. The underlying question is *how* interviewers fabricate, that is, which typical response patterns make them identifiable by statistical analyses. When discussing the issue, researchers commonly refer to the same response patterns in line with the assumption of a rationally acting falsifier. The empirical evidence on these response patterns, however, is far from conclusive. Results from different studies are often inconsistent and sometimes clearly contradictory.

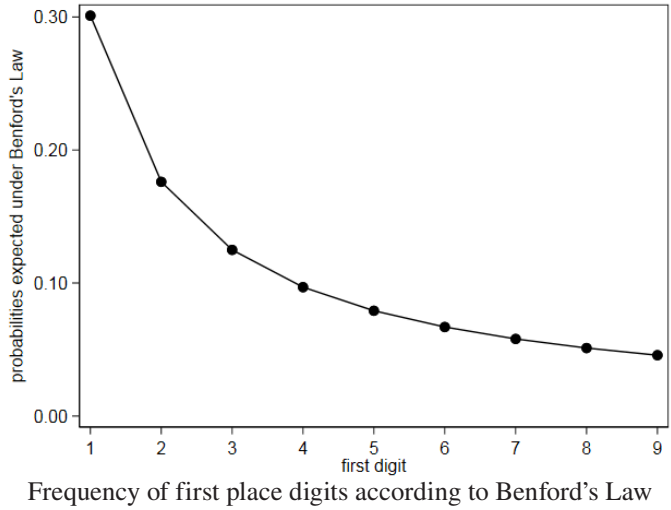
The response patterns that are typically expected from falsifiers will now be discussed in more detail, referring to previous studies that tested or used such patterns as indicators for fraud. The aim of the following section is twofold: It will introduce the internal indicators that will be empirically tested later on and it gives a first rough impression about how the indicators performed in fraud detection in the literature to date.

Nonconformity with Benford's Law (1)

Benford's Law implies that multi-digit numbers are more likely to start with a small than with a high digit (Benford 1938, Diekmann & Jann 2010). The frequency that a is the first digit in multi-digit numbers follows the distribution:

$$F_a = \log_{10}\left(\frac{a+1}{a}\right)$$

As Graph 1 shows, the first digit is one in approximately 30% of all numbers. With each next higher digit, the values continue to decrease.



Graph 1 Frequency of first place digits according to Benford's Law

What makes Benford's Law a potentially helpful tool for fraud detection is that it is perceived as "quite counterintuitive" (Nigrini 1999): Falsifiers should expect equal frequencies for all possible first digits in a number. Based on this assumption, Benford's Law has not only been used for fraud detection in survey data but also in regression coefficients of journal publications (Bauer & Gross 2011; Diekmann 2007). The extent to which values comply with Benford's Law is then assessed by a chi-squared test, measuring the difference between the observed and the expected distribution.

Various studies examine monetary values to identify falsifiers in the German Socio-Economic Panel (GSOEP): While Schäfer, Schräpler, Müller, & Wagner (2005) and Schräpler (2011) are mildly positive about their results, Schräpler & Wagner (2003) describe the method as "not efficient". Interestingly, they extend the analysis to the second digit distribution, which does not bring about convincing results. Also the true data contains far too many zeros due to rounding errors. A similar problem leads Porras and English (2004) to successfully test different variants of Benford's Law, including one that excludes the digit 5. Bredl et al. (2012) come to the conclusion that their falsifiers diverge more from Benford's Law than real respondents, although the difference is not very big. All in all, empirical evidence is mixed. Nonetheless, Benford's law is often simply assumed in scientific studies and used as an adequate means of detecting fraud, without verifying that, on the one hand, the real data follow the distribution and, on the other hand, that the falsifications do not (Diekmann & Jann 2010: 398).

Avoiding extreme categories (2)

It is usually assumed that falsifiers avoid the more obtrusive extreme categories in ordinal response scales, leading to smaller variances in faked data. Most empirical studies seem to find such differences in either the number of extreme categories or variances (e.g. Bredl et al. 2012; Kemper & Menold 2014; Schäfer et al. 2005). However, there are some exceptions: Menold & Kemper (2014) report inconsistent results and Schnell (1991) cannot find any effect, although he theoretically argues that falsifiers should generally underestimate the heterogeneity of respondents.

Strategic use of filter questions to shorten questionnaire (3)

An empirically rather uncontested hypothesis is that falsifiers make use of their knowledge about the filter branches in a questionnaire to find the shortest and easiest way through it (Menold & Kemper 2014; Brüderl, Huyer-May, & Schmiedeberg 2013; Josten & Trappmann 2016). In other words, it should be a promising strategy to take a closer look at the answers that interviewers gave to "gate questions" (Weinauer 2019), that is, questions that result in a list of follow-up questions. Alternatively, the number of inapplicable questions can be examined.

Avoiding item-nonresponse (4)

For the remaining questions that cannot be avoided by filter questions, contrastingly, it is believed that falsifiers provide answers more consistently throughout the questionnaire than real respondents because they do not want to raise suspicion. They hence should produce less item-nonresponse. Bredl et al. (2012) confirm this empirically, while Schnell (1991) finds the opposite effect.

Avoiding open answers (5)

Another common idea is that falsifiers should avoid open answers. This makes sense from a rational choice perspective: On the one hand because fictitious answers might be easily detected and on the other hand because it is comparatively burdensome and time-consuming to invent and write down a plausible answer. Empirically, things seem less clear: While Bredl et al. (2012) find fewer "other"-answers in their falsifications, Menold & Kemper (2014) come to opposite conclusions.

Hypotheses

After discussing these most common internal indicators, in the light of rational choice theory and with respect to prior findings from previous research, the subsequent part of this paper will subject the indicators to an empirical test. If the traditional assumptions of rational behavior hold, we generally expect to see the following response patterns in falsified data:
Compared to authentic respondents, …

1) … falsifiers violate Benford's Law when they report numbers.
2) … falsifiers choose fewer extreme categories on ordinal response scales.
3) … falsifiers use filter branches to shorten the questionnaire.
4) … falsifiers produce less item-nonresponse within their filter path.
5) … falsifiers give fewer open answers.

Trick question (6)

Apart from these rather commonly used indicators, we will also empirically test an unconventional approach to fraud detection, namely, a trick question which was deliberately designed and implemented as a potential method to identify fraud in this specific survey (for a different approach to trick questions where respondents are asked about fictitious words or newspapers, see Ziegler, Kemper, & Rammstedt 2013; or Menold & Kemper 2014; Winker, Kruse, Menold, & Landrock 2015 for implementations).

Respondents were asked for the European country with the highest income inequality. However, on their training, interviewers received false information about likely responses. As a consequence, we would expect falsifiers to avoid the presumably rare true answer "Portugal", while unsuspicious Eastern European countries should be mentioned more often than in real interviews.
These considerations result in the following additional hypotheses:

6a) … falsifiers avoid the presumably rare true answer "Portugal".

6b) … falsifiers overestimate the share of mentioned Eastern European countries.

More details on the trick question are provided in the section on data and methods.

Data and Methods

To test the potential of the discussed indicators for fraud detection, we use data from a cross-sectional survey on the fairness of earnings, in which authentic cases of fraud were detected during fieldwork. The survey was part of the project "The factorial survey as a method for measuring attitudes in population surveys", funded by the German Research Foundation (DFG). The questionnaire contained single-item and vignette questions on income-related fairness perceptions, some knowledge questions about income and labour in Germany, information on the respondent's own income, occupation and working environment, as well as questions on the respondent's socio-demographic background and a social desirability scale. All in all, the questionnaire was of moderate length: 70% of face-to-face respondents completed the questionnaire in 20 to 30 minutes.

The survey was conducted nationwide among the residential population of Germany aged 18+. In about 50% of cases, data were collected by interviewers in computer-assisted face-to-face interviews (CAPI). The sampling strategy comprised the random selection of 129 sample points throughout Germany, a random route procedure and a Kish–selection grid. The other half of respondents was recruited via telephone by means of random digit dialing in combination with a Kish-selection grid. This group completed a self-administered paper or online questionnaire (PAPI/CAWI). Since there was no possibility for interviewer falsification in the self-administered sample, the present project focuses on the 821 interviews conducted in the face-to-face setting. The 803 self-administered questionnaires are only occasionally mentioned for purposes of comparison.

The fieldwork monitoring comprised a range of external control mechanisms: In a first step, re-contact via postcard and paradata of interview time and duration were used to identify suspicious cases. In a second step, the suspicious cases were subjected to repeated contact attempts by telephone and follow-up checks of the random route. This process identified 44 falsifications. These falsifications were

admitted by the interviewers and consequentially deleted in consultation with the survey agency.

Fraud occurred in ten different sample points, seven of which were completely and three partially removed from the official data set. Since interviewer characteristics were not made available for the faked data, we need to make the (reasonable) assumption that each sample point was assigned to one interviewer in order to correct for clustered standard errors in the significance tests throughout the empirical analyses of this paper. If this assumption is true, the faked data were produced by ten different falsifiers. Out of these, seven interviewers falsified all of their work (five to seven interviews), while three only falsified one or two of their interviews.[1] To capture this pattern adequately and allow for the fact that interviewers have not necessarily faked all of their assigned interviews, the subsequent analyses will be carried out on the interview level. Looking at the existing literature on interviewer falsifications, this is a somewhat unusual approach. However, it best reflects the nature of our data and accounts for the fact that the number of conducted interviews per interviewer is too small to run reliable analyses at the interviewer level.

With regard to the explanatory variables, indicators are generally obtained by summing up over all available questions and separately for each interview. Put concretely, this is done for the first and second digits of the monetary values, the extreme categories on ordinal scales, inapplicable questions, missing values in mandatory questions, and the prevalence of open answers. We will use all of these internal indicators to compare real and fake interviews in order to obtain information on the extent to which falsifiers leave detectable traces within questionnaires. First, this will be done by descriptive comparisons between the two groups (page 13-20). The statistical tests reported for the indicators 2 to 6 are tests for mean comparisons, Somers'D or tests of proportions dependent on the variable's level of measurement. All of them use cluster-robust standard errors to account for the fact that interviews are nested within interviewers. In a second step, multivariate analyses are presented. In a logistic regression model, in which the dependent variable indicates whether an interview was actually conducted or fabricated, we will assess the relative importance of the explanatory variables and identify the most promising indicators for fraud detection. Again, cluster-robust standard errors are applied to account for the fact that observations are not independent but nested within interviewers.

1 Apart from one exception, falsifiers had a workload of five to seven interviews per interviewer. There is one falsifier with a workload of 13 interviews who falsified one of them. On average, unsuspicious interviewers completed more, that is, 12 interviews. For unsuspicious interviewers, numbers ranged from two to 28 interviews per interviewer.

Before moving on to the results, the remaining part of this subsection will provide more details on the concrete survey questions that the individual explanatory indicators rely on. The full question wording is provided in Appendix B.

- **Nonconformity with Benford's Law (1)**
  The questionnaire contained five monetary variables that can be checked for their compliance with Benford's Law: The estimated average monthly gross income for a full position in Germany, the respondent's own monthly gross income, the own gross income that the respondent would perceive as fair, the net household income per month and the household income necessary to pay for recurring expenses. For all of these items, respondents were asked to provide an open answer. While the available number of numeric values is too small to conduct analyses at the level of individual interviews or interviewers, comparing the real and the faked data as a whole can produce valuable insights concerning the usability of the Benford distribution for fraud detection.
- **Avoiding extreme categories (2)**
  To examine the shares of extreme response categories, we use two item batteries with 7-point Likert response scales. These questions were answered by all respondents, that is, they could not be inapplicable. In the first item battery, respondents were asked to evaluate to which extent certain person characteristics should have an impact on an employee's fair gross income. This task was completed for eleven different characteristics (e.g. sex and education) using an ordinal scale ranging from 0 ("not important at all") to 6 ("very important"). This item battery is followed by a social desirability scale, in which the respondents assessed their own personality on the basis of six statements that they evaluated on a scale from 0 ("not applicable at all") to 6 ("fully applicable"). Graph 3 shows the frequency of each response category for real and fabricated interviews across all 17 items.
- **Strategic use of filter questions to shorten questionnaire (3)**
  Twenty of the survey questions can in principle be skipped due to filters. A falsifier who knows the questionnaire could answer the filter questions strategically to shorten the questionnaire, resulting in higher numbers of inapplicable questions within filter paths.
- **Avoiding item-nonresponse (4)**
  There were 43 closed-ended questions that were mandatory for every respondent and could be subjected to an examination of item-nonresponse. The respective indicator consists of the sum of missing values in these questions.
- **Avoiding open answers (5)**
  There are seven open answers in the questionnaire that can be used to test hypothesis 5. Three of them are open questions in the stricter sense of the word, namely the job title and description of the respondent's current (or last) occupation, a feedback question at the end of the questionnaire and a trick ques-

tion on the European country with the highest income inequality (which we will come back to in in the results section). Apart from that, there were seven occasions in which respondents could specify additional options ("other, please specify:_____"). This was possible for their current occupation, their main place of residence since birth, the sources of their household income, their party preference, their partner's occupation, their highest educational degree, and their vocational qualifications. However, the latter three were not used by any respondent. We are hence left with seven potential open answers to analyze. For the descriptive analyses, the overall number of open answers serves as an indicator. For the regression model, we distinguish between completely open questions and text fields, where respondents could specify "other" options.

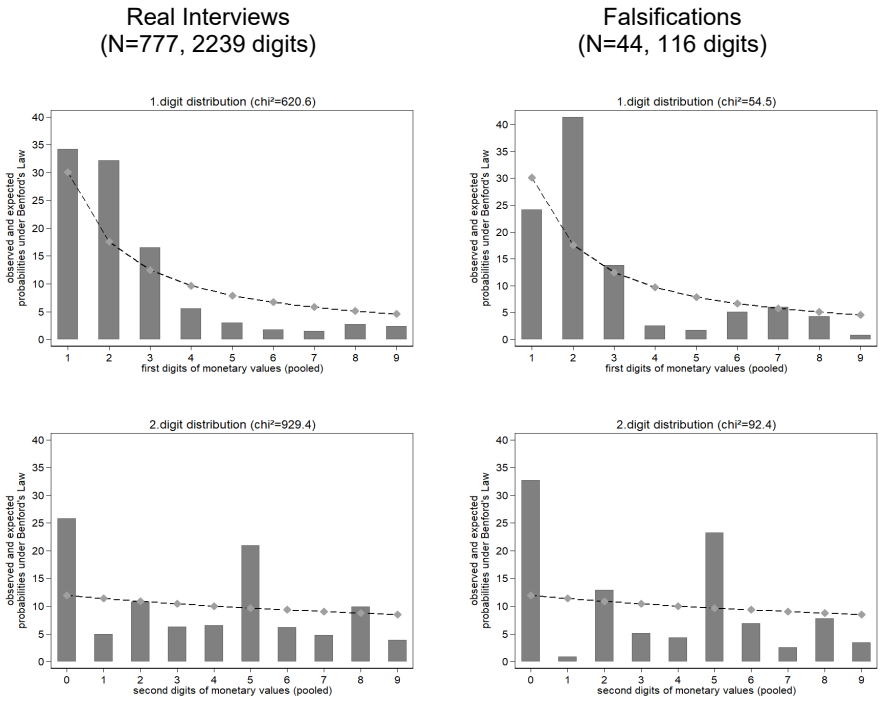
- **Trick question (6)**
  Implementing a trick question was an innovative attempt to detect fraudulent data in the setting of this specific survey. As part of an item battery that measured the respondents' general knowledge of the survey topic, they were asked for the European country with the highest income inequality. In the training that all the interviewers underwent before the fieldwork period, interviewers were informed about the right answer (which was Portugal at the time). However, they were also told that only experts know this and that respondents would usually guess an „Eastern European country such as Poland or Romania" (Sauer, Auspurg, Hinz, & Liebig 2010: 79). We will use the answer to the trick question to examine if such a trick question is helpful to detect fraud. This would be the case if falsifiers were more likely then real respondents to provide a presumably unsuspicious country, while avoiding the true but presumably rare answer.

Results

Benford's Law

Graph 2 compares the theoretically expected Benford distribution (curves) and the observed distribution (bars) separately for the first and the second digit and the real and the faked interviews (see Graph A1 in the appendix for the digit distributions in the self-administered survey modes).

A first obvious result is that there are considerable deviations from Benford's Law for certain digits in all four subgroups. Looking at the first digit of the real data, it is noticeable that the digits one to three are overrepresented compared to Benford's predictions, while the digits four to nine fall short of the expected percentages. Especially for numbers beginning with the digit two, the observed and expected values diverge strongly, with a difference of almost 15 percentage points. A look at the first digit distribution of the fabricated interviews reveals a similar



Graph 2 Compliance with Benford's Law

picture. The percentage of monetary values that begin with the digit two is even larger than in the real interviews: the observed value is 41% (and thus nine percentage points higher than in the case of the real interviews), while Benford's expectation would range just under 18%. The probability for an initial digit one, on the other hand, is six percentage points below the 30% predicted by Benford.

For the second digit distribution, we find clearly elevated shares of zeros and fives: In the real data, the probability of observing these digits is 14 and 11 percentage points higher than expected, in the fabricated data it is 21 and 14 percentage points above Benford's prediction. In line with the findings documented by some of the previous studies on Benford's second digit (Porras & English 2004; Schräpler & Wagner 2003; Winker et al. 2015), this is clear evidence for rounding. Interestingly, this tendency is slightly more pronounced in the fabricated interviews (although the difference is not statistically significant).

To sum up, neither the authentic nor the fraudulent data followed Benford's Law. Accordingly, chi-squared tests lead to a clear rejection of the hypothesis that the observed data follows Benford's distribution for the four subgroups.

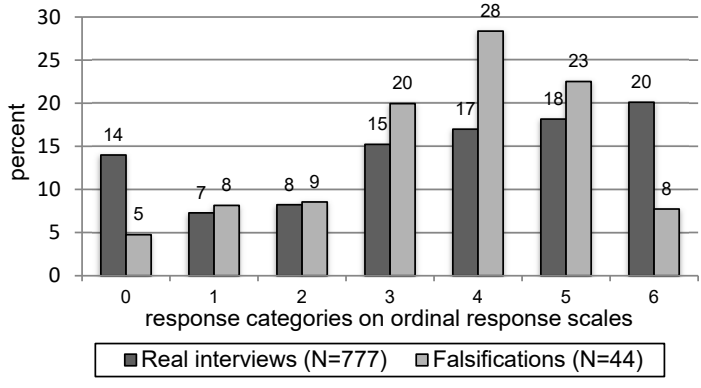
Judging from this finding, we would advise studies with a higher workload per interviewer to consider taking an alternative approach: Instead of comparing fal-

sifications to Benford's Law, each indiviudal interviewers' numerical values could be compared to the empirical sample distribution (as suggested by Swanson, Cho, & Eltinge 2003; Winker 2016). Considering the small number of interviews per interviewer in this data set, we will refrain from looking further into such analyses on the interviewer level.

Extreme Categories on ordinal response scales

As Graph 3 shows, the average number of extreme categories is much lower in the faked data than it is for real interviews. This is true both for the lowest response category 0, which real respondents choose in 14% and falsifiers in 5% of all cases, and for the highest response category 6, which real respondents choose in 20%, and falsifiers in 8% of their answers (both differences in proportions are statistically significant with $p<0.01$). This means that, while falsifiers underestimate the share of extreme categories by nine to twelve percentage points, the middle categories - and here in particular category 4 - are more frequently chosen in the faked data.[2]

This result is in line with what we find if we sum up the standard deviations within the ordinal response scales for each interview. The average standard deviation for the fake interviews is 1.51, for real interviews 1.94. Comparing the groups in a mean comparison test with cluster-robust standard errors, the difference is statistically significant ($p<0.001$).



Graph 3 Pooled responses to 17 questions with ordinal response scales

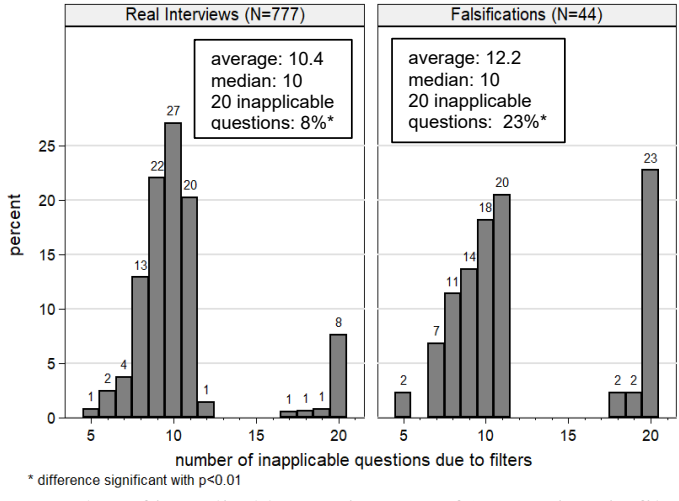
2 To see if this result was driven by very few falsifiers, the average number of extreme values was individually looked at and compared to the the average number of extreme categories in the real interviews (which was 5.7). Only one out of ten falsifiers (who moreover only faked one interview) ticked a slightly higher number of extreme categories (namely 6). All other falsifiers behaved in line with our hypothesis. Five of them ticked an average of only 2 or fewer extreme categories per interview.

Inapplicable questions in filter branches

We have argued that we would expect falsifiers to use filter branches that shorten the questionnaire. Graph 4 offers evidence that falsifiers indeed pursue this strategy. It shows the absolute number of inapplicable questions, separately for real and fabricated data.

At a first glance, the distributions look similar: The majority of cases have between five and twelve inapplicable questions in filter paths (90% for real interviews and 73% for faked interviews), while a comparatively smaller group avoids 17 or more questions. Differences are particularly evident in this upper part of the distribution. Compared to real interviewers, falsifiers are almost three times as likely to leave all 20 skippable questions unanswered (8% versus 23%; difference in proportions is statistically significant with $p<0.01$).[3]

The filter path with the largest number of follow-up questions that can be skipped is the one on employment history and working conditions. Interviews can only be in the upper group with 17 to 20 inapplicable questions if the response to the filter question is unanswered or indicates that the respondent has never been in employment. Falsifiers had a strong incentive to answer this question in the negative, because 16 questions on employment were omitted if the respondent had never worked.



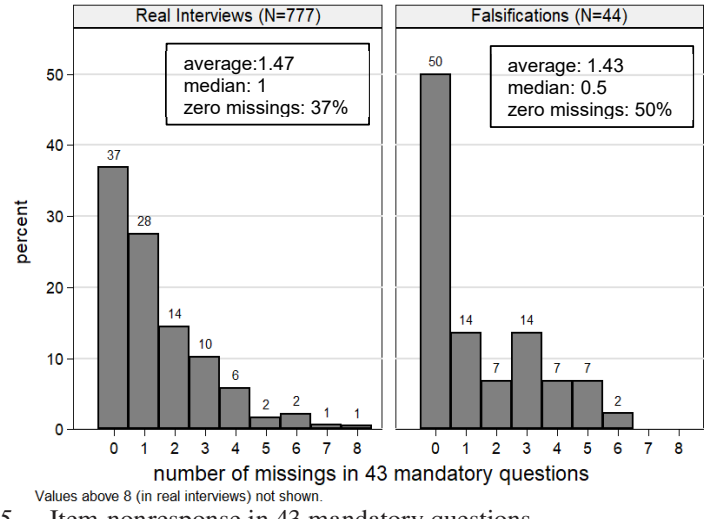
Graph 4 Number of inapplicable questions out of 20 questions in filter branch

3 As a robustness check, we individually compared each falsifier's average number of inapplicable questions to the average in real interviews to see if results could be driven by very few outliers. 4 out of 10 falsifiers' means ranged between 9.3 and 10, that is slightly lower than the average of 10.4 in real interviews. However, 6 out of 10 falsifiers showed the expected pattern and partly made extensive use of inapplicable questions to presumably shorten the questionnaire.

Item-Nonresponse

Based on the questions that could not be avoided by filters, Graph 5 compares the number of missings in the real and the faked interviews. Our theoretical argument was that falsifiers should avoid missing values in the questions they cannot skip, because too much item-nonresponse might attract the survey agency's attention.

In line with the expectations, falsifiers are more likely to produce questionnaires with zero missing values (50% versus 37%). However, this difference is not reflected in any significant differences, neither in the shares of zero missings nor in the group averages.[4]



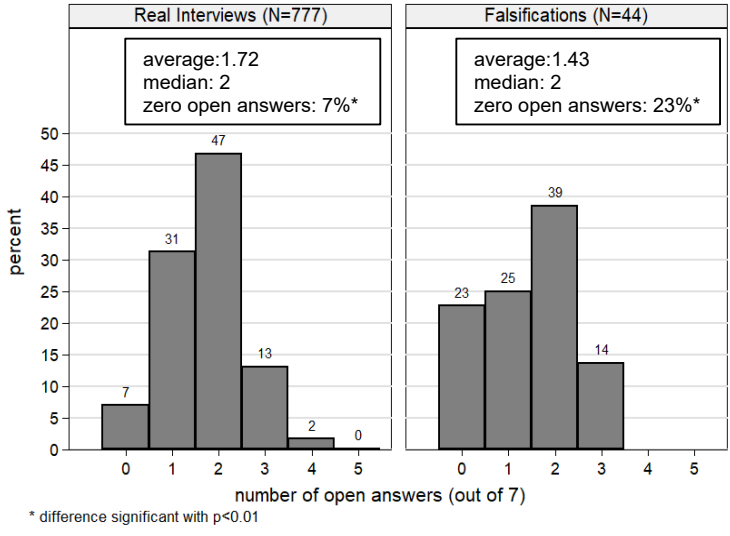
Graph 5 Item-nonresponse in 43 mandatory questions

4 When looking at the falsifiers' mean numbers of missings, 5 out of 10 yield numbers below the average of real interviews, 4 of them even below 0.2. The other half ranges between 2.2 and 5 missings per interview. These results hint towards heterogeneous interviewer behaviours.

Open answers

Graph 6 shows that there is no significant difference in the mean or median number of open answers between real and fabricated interviews. There is, however, a difference in whether open ended answers are given at all: 7% of real respondents do not give any open answers, compared to 23% of the falsifiers (difference in proportions is significant with $p<0.05$).[5]

Additional analyses that compared the number of letters in open answers (if there were any) across groups have not revealed any differences in means or medians (see Graph A2 in the appendix for a graphical representation).



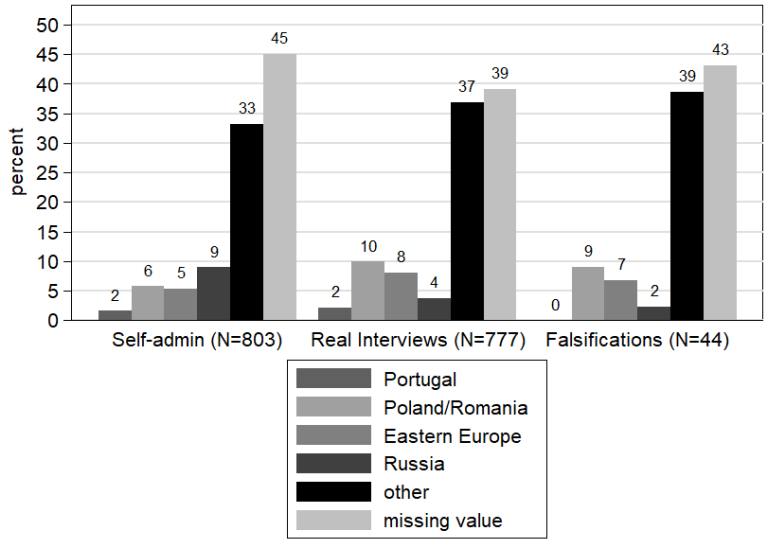
Graph 6 Open answers to seven open questions

5 7 out of 10 falsifiers showed lower average numbers of open answers compared to the average in real interviews. Comparing the prevalences of interviews with zero open answers, half of the falsifiers provided at least one open answer per interview. The other half had a likelihood of 29 to 100% to avoid open answers completely.

Trick question

Graph 7 shows the answers to the trick question. In addition to the real and the fabricated personal interviews, the answers given in the self-administered (paper and web) questionnaires are also reported to provide a reference point to a respondent group that could not have been affected by interviewer fraud.[6]

Interestingly and in line with hypothesis 6a, none of the falsified interviews mentioned the correct answer, Portugal. However, it was not a very frequent answer in the real interviews or in the self-administered questionnaires. The prevalence of Poland, Romania or other Eastern European countries did not differ enough (or in the expected direction) across groups to serve as an indicator for fraud in the face-to-face sample. We therefore reject hypothesis 6b.



Graph 7 Answers to trick question on income inequality in Europe across modes

6 Recruitment into the two survey modes took place by drawing two separate random samples of the residential population of Germany (more information in Sauer et al. 2010). For the purposes at hand, we are implicitly assuming that the random differences between samples do not affect the answers to the trick question.
As one of the reviewers pointed out, self-administered questionnaires might still contain falsifications if the survey agency added invented cases. Although this is a theoretical possibility, we have no reason whatsoever to believe that this happened in our data.

As expected, the answer Portugal predicts perfectly that an interview is authentic. Since the other answers do not help to distinguish between real and faked interviews, the trick question will not be further analyzed in the following multivariate analyses. It will, however, be included as one of the open questions, which can either be answered or missing.

Multivariate analyses

To examine the relative importance of the different indicators, the next step is to use them as potential predictors for fraud in a logistic regression model. The dichotomous dependent variable indicates whether an interview is real (coded as 0) or a case of known fraud (coded as 1).

Table 1 summarizes all explanatory variables. For the indicators on *extreme response categories* and *item-nonresponse*, some categories at the upper end of the scale were grouped together due to small numbers of cases.

Table 1 Explanatory variables in regression model

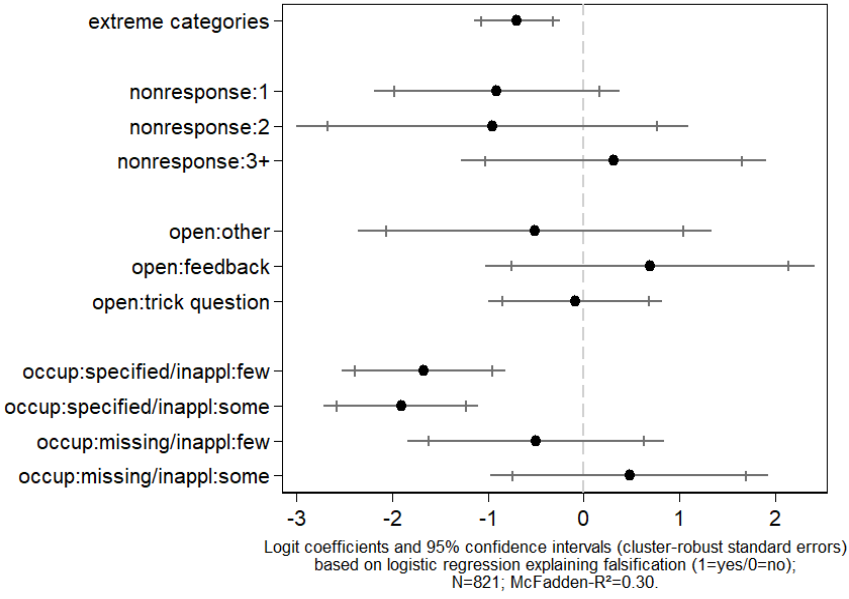
Variable	Concept and Coding	
extreme categories	number of extreme categories on 17 items with ordinal response scales (continuous variable)	→ hypothesis 2
item-nonresponse	number of missings in mandatory/unfiltered questions (0/1/2/3+, specified as dummies)	→ hypothesis 4
open: other	answered any "other, please specify" category (0/1)	→ hypothesis 5
open: feedback	answer to feedback question at the end of the survey (0/1)	→ hypothesis 5
open: trick question	answer to question on income inequality in Europe (0/1)	→ hypothesis 5
occupation * inapplicable	open question on occupation (skipped / specified / missing) number of inapplicable questions (5-9 "few"/ 10-12 "some"/ 16-20"many") 0: occupation: skipped / inapplicable: many 1: occupation: specified / inapplicable:few 2: occupation: specified / inapplicable: some 3: occupation: missing / inapplicable: few 4: occupation: missing / inapplicable: some	→ hypothesis 5 → hypothesis 3

There is a peculiarity in the data concerning the open question that elicits the respondent's current (or alternatively last) *occupation*. It is positioned within the longest filter branch of the questionnaire that was designed to collect information on the respondent's employment and working conditions. Failure to specify an occupation can thus stem from item-nonresponse or from skipping the entire filter branch. A high number of *inapplicable questions* essentially indicates that the complete filter branch on employment, including the item on occupation, was skipped. In terms of coding, this correlational pattern is reflected in the interaction of two categorical variables, the respondent's reaction to the occupation question (skipped / missing / specified) and the number of inapplicable questions (few / some / many). The reference category indicates an inapplicable occupation question and a generally high number of inapplicable questions due to filters throughout the questionnaire.

Apart from respondent occupation, all other *open answers* are captured by dichotomous variables. Since very few respondents specified additional information in "other, please specify" categories, these questions were subsumed into one dummy variable indicating if any (versus no) open answer was provided.

Graph 8 shows the results of the logistic regression model (see Table A1 in the appendix for the full regression table). A first result is that there is a significant negative correlation between the number of extreme categories in the questionnaire and the log-odds that an interview was fabricated ($p<0.01$). This means that hypothesis 2 is clearly confirmed and falsifiers indeed avoided extreme answers on ordinal response scales. Equally clearly, we cannot find any evidence that item-nonresponse is related to fraud (hypothesis 4). Regarding hypothesis 5, it is interesting to note that the absence of open questions (or the numbers of letters in open answers as mentioned before) does not generally come with higher probabilities of falsification. However, if we look at the question on respondent occupation, questionnaires that contained a job description were significantly less likely to be falsifications than interviews with item-nonresponse ($p=0.08$ for *few* inapplicable questions, $p>0.01$ for *some* inapplicable questions) or a skipped question ($p<0.01$ for *few* and *some* inapplicable questions).

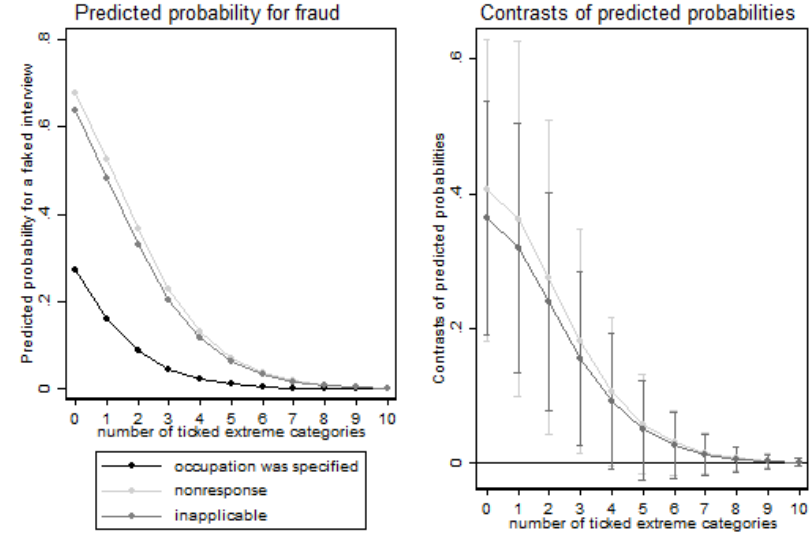
When it comes to inapplicable questions (hypothesis 3), regression coefficients for respondents that got filtered over *few* or *some* questions were very similar. Only a high number of skipped questions (including the occupation question) was predictive for fraud in our data. This pattern somewhat suggests that the intention to skip the open question on occupation was the dominant one compared to finding the shortest path through the filter paths of the questionnaire.



Graph 8 Potential Indicators for Fraud (logistic regression)

Looking into descriptives, falsifiers report around three times as many respondents who have never been in employment and thus automatically skip the 16 additional questions on employment (27.3 versus 9.5%). Among those who are asked about their current or previous occupation, 93.0% of the real interviews provide a job title, while only 81.3% of the falsifications do. With these considerable differences in response patterns, the question on the respondent's occupation was very helpful in identifying fabricated interviews. Part of the explanation surely lies in how easy it would be to check the correctness of the answer. Respondents certainly remember what they do or did for work, and even other household members could presumably provide the right answer, e.g. in a follow-up check by phone. However, the question is not only easily verifiable but also easily avoidable by a strategic use of the filter paths. The combination of both peculiarities seems to explain its success in identifying fraud.

Graph 9 translates the two significant indicators from the regression model into a visual illustration: It shows how the number of extreme categories and the



Graph 9 Predicted probability for fraud dependent on the strongest predictors

question on occupation interact.[7] The y-axis of the graph on the left side shows the predicted probability that an interview has been fabricated.

At a first glance, the graph shows that the effect of one variable heavily depends on the other variable. Generally speaking, the probability of fraud decreases with an increasing number of selected extreme categories on the ordinal response scales of the questionnaire. This is even more so if the open question on the respondent's occupation has not been answered - due to nonresponse or because the respective filter branch was skipped entirely.

Within the groups of interviews that did not indicate any occupation, the predicted probability of fraud can rise to a maximum of 63% or 68% if zero extreme categories have been ticked. For interviews in which the occupation was specified, the slope is less steep and only climbs to 27% if we move towards the lower end of the x-axis indicating low numbers of extreme categories. Comparing the interviews with and without the open answer, the predicted probability of fraud hardly

7 Graph 9 only refers to occupation rather than the interaction between occupation and inapplicable questions. This is justified by the fact that there weren't any significant differences between the groups with few and some inapplicable questions in the previous model. In addition, the number of cases with a missing value for the occupation question seemed too small (N=55) to identify further subgroups (as can be seen by the large confidence intervals for the last two regression coefficients in Graph 8).

changes at the upper end of the scale, while the open answer accounts for a change of 36 to 41 percentage points if no extreme category was chosen.

The illustration on the right side of Graph 9 treats interviews with a specified occupation as the reference category (straight line) and compares it to the two groups without valid answers (curves). The confidence intervals indicate that the group differences between the interviews with and without a specified occupation are significant if zero to three extreme categories were ticked and insignificant if four or more extreme categories were ticked.

Judging by the maximum changes in predicted probabilities that can be attributed to the variables, the results suggest that the number of extreme categories is an even stronger indicator of fraud than failing to provide a job title in a filter branch that falsifiers might want to skip.

Related findings: standard deviations, acquiescence and social desirability

Theoretically, we could have considered standard deviations in the ordinal response scales or the amount of straight lining (as done by Blasius & Thiessen 2018) instead of the number of extreme categories. Empirically, however, the number of ticked extreme categories was more strongly correlated with fraud and the standard deviations did not add any explanatory power to the regression models once the model controlled for extreme categories.

Following Kemper & Menold (2014), we tested two more alternatives to examining the extreme categories, namely acquiescence (in both ordinal response scales) and socially desirable answers (in the social desirability scale). In their paper on laboratory falsifications, Kemper & Menold (2014) report that falsifiers provide "overly positive self-descriptions" (p. 96) when asked about socially desirable behaviors as well as a higher tendency to acquiesce irrespective of question content. These results were not replicable with our data. Both options performed worse as indicators for fraud than the number of extreme categories. Generally, adding indicators for self- and other-deception did not help the predictive power of the regression model. Descriptively, falsifiers tended to give negative self-descriptions in five out of six cases. Curiously, they were less likely to agree to the statement "I am always honest to others". Although the difference between groups was not statistically significant, this finding would be in line with the argument that falsifiers use themselves as a reference point when fabricating data (see Landrock 2017).

Predicting fraud based on internal indicators alone?

In response to a reviewer comment, Table 2 presents the numbers of falsifications that would have been correctly classified. In other words, we are treating our data as a case of supervised learning relying on internal indicators (on interview level)

Table 2 Confusion Matrix

	Actual: Real Interview	Actual: Falsification	Total
Prediction: real	757	22	779
Prediction: falsified	20	22	42
Total	777	44	821

alone. As the analyses are carried out on the interview level, we are allowing for the situation that interviewers only faked some of their assigned interviews.

To predict fraud based on the logistic regression model, it makes sense to work with an educated guess about the expected share of falsifications instead of using the 50% mark as a cut-off point. Similar to ordinary clustering approaches, the method would otherwise be very likely to identify an unrealistically high number of falsifications (De Haas & Winter 2016). In this case, the upper 5% of interviews that were most suspicious from internal indicators are treated as falsifications.

The results suggest that only by taking internal indicators at interview level into account, we could have identified 22 out of 42 confirmed falsifications correctly. 20 would have remained undetected and there would have been 22 new suspicious cases (which could have been subjected to further checks if the internal indicators had been part of the quality control during fieldwork).

This little exercise highlights a point that has been made before: Internal indicators seem to work well to identify some falsifications but not others (a finding that is confirmed by Thissen & Myers 2016). As a consequence, it is not reasonable to rely only on one type of indicator. Instead it should be the goal to combine as many as possible. In settings where the number of conducted interviews per interviewer is higher, internal indicators on interviewer level can supplement the internal indicators on interview level that we used in this study. These could include duplicate checks (Koczela, Furlong, McCarthy, & Mushtag 2015), analysis of similar response patterns in ordinal response scales (Blasius & Thiessen 2015), and comparisons of so-called "content-related patterns", contrasting interviewer-specific and overall sample means (Kosyakova, Olbrich, Sakshaug, & Schwanhäuser 2019; Weinauer 2019).

To some extent (although in varying degrees), all internal indicators have the disadvantage that falsifiers will be able to adjust their response style if they are aware that a certain kind of check is in place (Winker 2016). However, the idea is not that one method will identify all falsifications, but that more checks will heigthen the chance of being detected and will make fraud less attractive to inter-

viewers. As Thissen & Myers (2016) put it: "Each method can be circumvented, but a combination of methods acts as a series of barriers, and patterns of falsification that might slip past one type of review may be caught by another."

Conclusion

This study provides empirical insights into the response patterns of falsifiers. It contributes to a wider body of literature aiming to develop more efficient monitoring strategies to prevent interviewer fraud and potential bias in surveys. For this purpose, we examined the potential of statistical ex-post analysis of response patterns for the identification of fabricated data. In contrast to the wide majority of studies on interviewer fraud, we did not draw on laboratory falsifications but could rely on authentic cases of interviewer fraud stemming from a survey project on income inequality in Germany. Out of 821 face-to-face interviews, 44 were identified as falsifications by external control mechanisms and admitted as such by the survey agency. These data were particularly suitable to test the potential of internal indicators for fraud detection: To what extent do the attempts of falsifiers to produce unsuspicious data lead to recognizable response patterns?

Drawing on somewhat contradictory empirical evidence from previous studies, hypotheses were formulated as to how rationally acting falsifiers would navigate through a questionnaire. Real and faked interviews were then compared with regard to various testable criteria, namely the number of selected extreme categories in ordinal response scales, answers to open-ended questions, item-nonresponse, strategic use of filter questions to shorten the questionnaire, and the compliance of reported numbers with Benford's Law. In addition, we reported results from an innovative trick question.

In the multivariate analysis, the avoidance of extreme categories on ordinal response scales proved to be the strongest indicator for fraud. This approach also outperformed alternative indicators, namely socially desirable answers and acquiescence in the ordinal response scale. Apart from that, missing data in one particular open-ended question significantly predicted fraud: the open question on the respondent's current or previous occupation. Missing values could occur either because the respondents refused to provide a job title or because they had never been in employment, in which case the interviewer was supposed to skip a filter branch of 16 questions on employment. None of the other open-ended questions, however, helped to detect fraud, and neither did the number of letters in open answers. This means that fewer or shorter open answers per se do not seem to be suitable indicators. Similarly, questionnaires with few and moderate numbers of inapplicable questions did not differ in their probability of being falsified. Only a high number of inapplicable questions (stemming from skipping the entire filter

branch on employment) was strongly correlated with fraudulent data. Although we cannot fully disentangle the effect of not answering an easily verifiable open-ended question and using filters to shorten the questionnaire, we can definitely say that the combination of both successfully identified fraud in our case.

Another interesting finding of the above analyses is that, contrary to the original hypothesis, falsifiers and real respondents did not differ in their shares of item-nonresponse. Apart from that, neither the implemented trick question nor Benford's Law were helpful in detecting fraud. Although all falsifiers avoided the rare correct answer to the trick question, very few real respondents gave the correct answer. When comparing the first and the second digits of reported monetary values to Benford's Law, this criterion proved to be highly problematic, since even the basic assumption that the authentic interviews should approximately follow the Benford distribution was violated. Judging from this finding, it seems more promising to experiment with rounding behavior (although differences failed to reach a significant level in our data) or to compare individual clusters of interviews to the average empirical distribution (as suggested by Swanson, Cho, & Eltinge 2003; Winker 2016). This approach can be pursued if the number of interviews per interviewer is sufficiently high to allow analyses at the interviewer level.

A limitation of this study is that we cannot rule out that, despite extensive checks, further unnoticed falsifications have remained in the data. Comparing the answers to the trick question and the digits from the monetary values to the questionnaires from the self-administered survey modes shows that this might be the case, although differences between real face-to-face interviews and self-administered questionnaires could in principal also stem from imperfect randomisation, mode differences in non-response bias or response behavior. Despite these uncertainties, the present study could show that more than half of our authentic cases of fraud which were detected by external control mechanisms would also have raised suspicion in a statistical ex-post analysis of their response patterns. This result can encourage researchers and survey agencies to use both types of indicators in combination when identifying suspicious cases.

One possible practical approach could be to run checks on response patterns more continuously during the fieldwork period. In line with AAPOR recommendations (AAPOR 2003), it is desirable to complement random picking of interviews with a more targeted strategy for additional quality control. This is where statistical approaches could help to identify suspicious cases for further checks. Ideally such a routine would already be at work during fieldwork, not after its completion. A corresponding statistical routine would have to identify outliers based on a set of internal indicators. This could be done by assigning suspicion points to interviews, by generating pareto charts of the collected paradata on a weekly basis (as suggested by Gwartney 2013). Concrete advise for implementing statistical routines that identifiy suspicious interviewers who deviate significantly from the overall

sample mean has been provided by Weinauer (2019) and Kosyakova et al. (2019) who draw on the concept of "meta-indicators".

Regarding the selection of internal indicators for statistical procedures, the presented results suggest that it is worth looking at the peculiarities of the specific questionnaire to e.g. identify high-risk questions that are easily verifiable or crucial for long filter paths. A general recommendation is to make use of as many indicators as possible - internal and external ones - to identify and substantiate suspicions.

References

AAPOR (2003). Interviewer Falsification in Survey Research: Current Best Methods for Prevention, Detection and Repair of Its Effects. American Association for Public Opinion Research (AAPOR). https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf

Bauer, J., & Gross, J. (2011). Difficulties Detecting Fraud? The Use of Benford's Law on Regression Tables. *Jahrbücher für Nationalökonomie und Statistik*, 231(5+6).

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society,* 78(4), 551–572.

Blasius, J., & Thiessen, V. (2018). Perceived Corruption, Trust, and Interviewer Behavior in 26 European Countries. *Sociological Methods & Research* (online first).

Blasius, J., & Thiessen, V. (2015). Should We Trust Survey Data? Assessing Response Simplification and Data Fabrication. *Social Science Research* 52, 479–493.

Blasius, J., & Thiessen, V. (2013). Detecting Poorly Conducted Interviews. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 67-88). Peter Lang Academic Research.

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

Brüderl, J., Huyer-May, B., & Schmiedeberg, C. (2013). Interviewer Behavior and the Quality of Social Network Data. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 147–160). Peter Lang Academic Research.

Bushery, J. M., Reichert, J. W., Albright, K. A., & Rossiter, J. C. (1999). Using date and time stamps to detect interviewer falsification. *Proceedings of the American Statistical Association*, 316–320.

Crespi, L. (1945). The Cheater Problem in Polling. *Public Opinion Quarterly* 9, 431-445.

De Haas, S., & Winker, P. (2016). Detecting Fraudulent Interviewers by Improved Clustering Methods – The Case of Falsifications of Answers to Parts of a Questionnaire. *Journal of* Official Statistics, 32(3), 643–660.

De Haas, S., & Winker, P. (2014). Identification of Partial Falsifications in Survey Data. *Statistical Journal of the IAOS* 30(3), 271–281.

Diekmann, A. (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. *Journal of Applied Statistics*, 34(3), 321–329.

Diekmann, A., & Jann, B. (2010): Benford's Law and Fraud Detection. Facts and Legends. *German Economic Review* 11(3), 397–401.

Dorroch, H. (1994). *Meinungsmacher-Report. Wie Umfrageergebnisse entstehen.* Göttingen: Steidl.

Esser, H. (1999). *Die Wert-Erwartungstheorie. Soziologie - Spezielle Grundlagen, Band I: Situationslogik und Handeln.* Frankfurt am Main: Campus Verlag, Kapitel 7.

Finn, A., & Ranchhod, V. (2017). Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South African Survey. *The World Bank Economic Review,* 31(1), 129–157.

Gwartney, Patricia A. (2013). Mischief versus Mistakes: Motivating Interviewers to Not Deviate. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 195–215). Peter Lang Academic Research.

Hood, C., & Bushery, J. (1997). Getting more bang from the reinterviewer buck: Identifying "at risk" interviewers. *Proceedings of the American Statistical Association*, 820–824.

Josten, M., & Trappmann, M. (2016). Interviewer Effects on a Network-Size Filter Question. *Journal of Official Statistics,* 32(2), 349–373.

Kemper, C. J., & Menold, N. (2014). Nuisance or remedy? The utility of stylistic responding as an indicator of data fabrication in surveys. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* 10(3), 92–99.

Koch, A. (1995). Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim ALLBUS 1994. *ZUMA Nachrichten,* 36, 89–105. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-208984

Koczela, S., Furlong, C., McCarthy, J., & Mushtaq, A. (2015). Curbstoning and beyond: Confronting Data Fabrication in Survey Research. *Statistical Journal of the IAOS* 31(3), 413–422.

Kosyakova, Y., Olbrich, L., Sakshaug, J., & Schwanhäuser, S. (2019). Identification of Interviewer Falsification in the IAB-BAMF-SOEP Survey of Refugees in Germany. FDZ-Methodenreport 02.2019. http://doku.iab.de/fdz/reporte/2019/MR_02-19_EN.pdf

Kroneberg, C., & Kalter, F. (2012). Rational Choice Theory and Empirical Research: Methodological and Theoretical Contributions in Europe. *Annual Review of Sociology,* 38(1): 73–92.

Landrock, U. (2017). How Interviewer Effects Differ in Real and Falsified Survey Data: Using Multilevel Analysis to Identify Interviewer Falsifications. *methods, data, analyses,* 11(2), 163-188.

Menold, N., & Kemper, C. J. (2014). 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,* 26(1), 41–65.

Menold, N., Landrock, U., Winker, P., Pellner, N., and Kemper, C. J. (2018). The Impact of Payment and Respondents' Participation on Interviewers' Accuracy in Face-to-Face Surveys: Investigations from a Field Experiment. *Field Methods*, 30(4), 295–311.

Menold, N., Winker, P., Storfinger, N., & Kemper, C. (2013). A Method for Ex-Post Identification of Falsifications in Survey Data. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 25–47). Peter Lang Academic Research.

Murphy, J., Biemer, P., Stringer, C., Thissen, R., Day, O., & Hsieh, Y. P. (2016). Interviewer falsification: Current and best practices for prevention, detection, and mitigation. *Statistical Journal of the IAOS* 32, 313–326.

Nelson, J. E., & Kiecker, P. L. (1996). Marketing Research Interviewers and Their Perceived Necessity of Moral Compromise. *Journal of Business Ethics* 15, 1107–1117.

Nigrini, M. (1999). I've got your number. *Journal of Accountancy* 187(5), 79–83.

Porras, J., & English, N. (2004). Data-Driven Approaches to Identifying Interviewer Data Falsification: The Case of Health Surveys. *Proceedings of the American Statistical Association*, 4223–4228.

Reuband, K.-H. (1990). Interviews, die keine sind. „Erfolge" und „Misserfolge" beim Fälschen von Interviews. *Kölner Zeitschrift Für Soziologie Und Sozialpsychologie: KZfSS,* 42, 706–733.

Sauer, C., Auspurg, K., Hinz, T., & Liebig, S. (2010). Konzeption und Durchführung der Studie „Einkommensgerechtigkeit in Deutschland" im Rahmen des Projekts „Der faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen". Feldbericht.

Schäfer, C., Schräpler, J.-P., Müller, K.-R., & Wagner, G. G. (2005). Automatic Identification of Faked and Fraudulent Interviews in the German SOEP. *Schmollers Jahrbuch: Journal of Applied Social Science Studies,* 125(1), 183–193.

Schnell, R. (1991). Der Einfluss gefälschter Interviews auf Survey-Ergebnisse. *Zeitschrift für Soziologie,* 20(1), 25–35.

Schräpler, J.-P., & Wagner, G. 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 No. 969. http://ftp.iza.org/dp969.pdf

Schräpler, J-P. (2011). Benford's Law as an Instrument for Fraud Detection in Surveys Using the Data of the Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik* 231.

Schreiner, I., Pennie, K., & Newbrough, J. (1988). Interviewer Falsification in Census Bureau Surveys. *Proceedings of the American Statistical Association*, 491–496.

Schwanhäuser, S., Sakshaug, J., Kosyakova, Y., & Kreuter, F. (2020). Statistical Identifcation of Fraudulent Interviews in Surveys: Improving Interviewer Controls. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter & B. T. West (Eds.), *Interviewer Effects from a Total Survey Error Perspective* (pp. 91-105). Chapman and Hall/CRC.

Storfinger, N., & Winker, P. (2013). Assessing the Performance of Clustering Methods. In P. Winker, N. Menold & R. Porst (Eds.), *Interviewers' Deviations in Surveys. Impact, Reasons, Detection and Prevention* (pp. 49-65). Peter Lang Academic Research.

Swanson, D. Cho, M, J., & Eltinge, J. (2003). Detecting Possibly Fraudulent or Error-Prone Survey Data Using Benford's Law. *Proceedings of the American Statistical Association*, 4172–4177.

Thissen, R. (2014). Computer Audio-Recorded Interviewing as a Tool for Survey Research. *Social Science Computer Review* 32(1), 90–104.

Thissen, M. R., & Myers, S. K. (2016). Systems and Processes for Detecting Interviewer Falsification and Assuring Data Collection Quality. *Statistical Journal of the IAOS* 32(3), 339–47.

Wagner, J., Olson, K., & Edgar, M. (2017). The Utility of GPS Data in Assessing Interviewer Travel Behavior and Errors in Level-of-Effort Paradata. *Survey Research Methods* 11, 218-233.

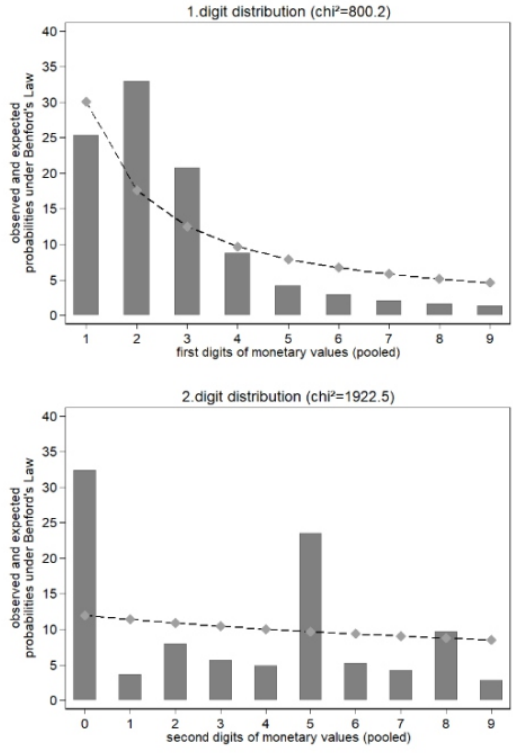
Weinauer, M. (2019). Be a Detective for a Day: How to Detect Falsified Interviews with Statistics. *Statistical Journal of the IAOS* 35(4), 569–75.

Winker, P. (2016). Assuring the Quality of Survey Data: Incentives, Detection and Documentation of Deviant Behavior. *Statistical Journal of the IAOS 32*(3), 295–303.

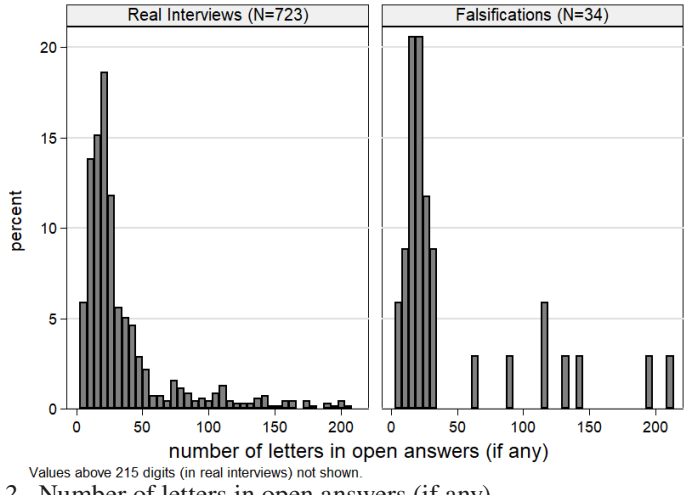
Winker, P., Kruse, K.-W., Menold, N., & Landrock, U. (2015). Interviewer Effects in Real and Falsified Interviews: Results from a Large Scale Experiment. Statistical Journal of the IAOS 31(3), 423–434.

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The Vocabulary and Overclaiming Test (VOC-T). Journal of Individual Differences 34(1), 32–40.

Appendix A



Graph A1 Compliance with Benford's Law in the self-administered surveys (N=803, 2612 digits)



Graph A2 Number of letters in open answers (if any)

Table A1 Potential indicators for fraud – regression table

Logistic Regression explaining falsification (1=yes/0=no) (logit coefficients; cluster-robust standard errors)	
extreme categories	-0.70** (-3.07)
nonresponse:1	-0.91 (-1.39)
nonresponse:2	-0.96 (-0.92)
nonresponse:3+	0.31 (0.38)
open:other	-0.51 (-0.54)
open:feedback	0.69 (0.79)
open:trick question	-0.09 (-0.19)
occup:specified/inappl:few	-1.67*** (-3.83)
occup:specified/inappl:some	-1.91*** (-4.63)
occup:missing/inappl:few	-0.50 (-0.73)
occup:missing/inappl:some	0.48 (0.64)
intercept	1.05 (1.67)
McFadden Pseudo-R^2	0.31
N	821

t statistics in parentheses

* p<0.05, ** p<0.01, *** p<0.001

Appendix B. Question Wording by Indicators

(for the original questionnaire in German language, see Sauer et al. 2010)

- **Nonconformity with Benford's Law**

What do you think, what is the average gross income per month for a full time position in Germany?
(This is the amount that employees receive from their employer before taxes and social security contributions are deducted.)
About _____ ,- Euro per month

What is your own monthly gross income from your employment?
(This is the amount that you receive from your employer before taxes and social security contributions are deducted. If you are self-employed, please fill in the average of what you earn per month.)
About _____ ,- Euro per month

In case you don't perceive your own income as fair, what would be a fair monthly gross income for you?
About _____ ,- Euro per month

What is the monthly net income of your household overall?
(Please sum up all types of household income, including your own. This also includes income from rentals and royalties, pensions, unemployment, social security and child benefits, rent subsidy and other types of income.)
About _____ ,- Euro per month

To cover for your running expenses, what is the minimum monthly net income that your household would need?
(Please specify the amount you need per month to cover for housing, food, clothes, heating and your personal basic needs.)
About _____ ,- Euro per month

- **Avoiding extreme categories**

In your opinion, to what extent should the following factors matter for a fair gross income?

Age …….. not at all 0-1-2-3-4-5-6 very much
Sex …….. not at all 0-1-2-3-4-5-6 very much
Education …… not at all 0-1-2-3-4-5-6 very much
Number of children …….. not at all 0-1-2-3-4-5-6 very much
Job ……... not at all 0-1-2-3-4-5-6 very much
Job experience …… not at all 0-1-2-3-4-5-6 very much
Health condition ……... not at all 0-1-2-3-4-5-6 very much
Time working for company …… not at all 0-1-2-3-4-5-6 very much
Size of company …….. not at all 0-1-2-3-4-5-6 very much
Economic situation of the company not at all 0-1-2-3-4-5-6 very much
Performance on the job ……... not at all 0-1-2-3-4-5-6 very much

To what extent do you agree with the following statements?

The first impression I have of people usually
turns out to be right..…… not at all 0-1-2-3-4-5-6 very much
I am usually very sure of my judgements..…… not at all 0-1-2-3-4-5-6 very much
I am not always aware of the reasons for my
actions..…….. not at all 0-1-2-3-4-5-6 very much
It has happened that I kept too much change..... not at all 0-1-2-3-4-5-6 very much
I am always honest with other people…….. not at all 0-1-2-3-4-5-6 very much
I have never taken advantage of somebody…..... not at all 0-1-2-3-4-5-6 very much

- **Avoiding open answers**

What is your current job, or what was your last job? Please specify the job title and describe your role exactly.

What do you think, which European country does currently have the highest income inequality?

We might have missed something that you consider important. Is there anything you want to add or comment on?

methods, data, analyses | 2021, Vol. 15(2), pp. 161-190 DOI: 10.12758/mda.2021.05

In Search of the Best Response Scale in a Mixed-mode Survey (Web and Mail). Evidence from MTMM Experiments in the GESIS Panel

Hannah Schwarz[1], *Wiebke Weber* [1], *Isabella Minderop*[2] *& Bernd Weiß*[2]

[1] *RECSM, Universitat Pompeu Fabra*

[2] *GESIS – Leibniz Institute for the Social Sciences*

Abstract

Mixed-mode surveys allow researchers to combine the advantages of multiple modes, for example, the low cost of the web mode with the higher coverage of offline modes. One drawback of combining modes is that there might be systematic differences in measurement across modes. Thus, it would be useful to know which measurement methods work best in all employed modes. This study sets out to find a method that results in the highest measurement quality across self-administered web mode questionnaires (web mode) and self-administered paper questionnaires sent out by mail (mail mode). Two Multitrait-Multimethod (MTMM) experiments employing questions on environmental attitudes and supernatural beliefs were implemented in the GESIS Panel, a probability-based panel in Germany. The experiments were designed to estimate the measurement quality of three different response scales: A seven-point fully labelled scale, a 101-point numerical open-ended scale and an eleven-point partially labelled scale. Our results show that the eleven-point partially labelled scale consistently leads to the highest measurement quality across both modes. We thus recommend using eleven-point partially labelled scales when measuring attitudes or beliefs in mixed-mode surveys combining web and mail mode.

Keywords: measurement quality; length of response scales; labelling of response scales; Multitrait-Multimethod (MTMM); mixed-mode; self-completion; web mode; mail mode

While in the past surveys were mainly unimode, nowadays respondents often receive the possibility to answer in the mode of their choice. This is supposed to increase their willingness to participate and lower survey costs (Eifler & Faulbaum, 2017). Mixed-mode surveys are used in various settings, especially where certain population groups are difficult to reach via the main survey mode. For example, these surveys are useful when researchers aim to conduct web mode surveys but have to account for the fact that parts of the target population do not use the internet (Bosnjak et al., 2017; ESOMAR & WAPOR, 2014). While this is an adequate strategy to deal with coverage error, it may lead to issues concerning measurement equivalence, as respondents may answer questions differently across modes (ESOMAR & WAPOR, 2014; Grewenig et al., 2018; Blom et al., 2016). Linked to this, employing the same measurement instrument across different modes can also lead to differences in measurement quality (e.g., Sánchez Tomé, 2018; Tourangeau, 2017; Dillman et al., 2014).

Measurement quality, in general, refers to the relationship between the unobserved, latent variable of interest and the observed response, and is here defined as the product of validity and reliability (Saris & Andrews, 1991). Validity here covers the construct validity subtypes convergent and discriminant validity (Campbell & Fiske, 1959). More specifically, it is defined as the strength of the relationship between a latent variable of interest and a so-called 'true score'. This 'true score' represents the score that respondents would have provided if no random measurement error existed. Reliability, in the model we employ, is defined as the strength of the relationship between the 'true score' and the observed variable. It captures the absence of random measurement error. It should be noted that an array of different definitions and operationalizations of measurement quality, as well as of validity and reliability, are used in the literature (Saris & Andrews, 1991). While studies may differ along these lines, they share the aim of empirically capturing measurement quality, that is, the absence of measurement error. The above definition of measurement quality thus holds for our analysis, while a somewhat broader perspective on the concept will be considered in the literature review.

Questionnaire designers in mixed-mode settings need not only to make sure that they indeed measure what they aim to measure. Furthermore, they need to ensure that respondents in different modes understand the measurement instru-

Acknowledgements

We thank Sonja Paulick-Fabini as well as the anonymous reviewer for their helpful comments.

Direct correspondence to

Hannah Schwarz, RECSM, Universitat Pompeu Fabra,
Sociometric Research Foundation
E-mail: hannah.schwarz@upf.edu

ment similarly. This means that questionnaire designers employing multiple modes have to make decisions bearing in mind the various features of the different survey modes. In self-administered modes, respondents can always see the questions and response scales, while in interviewer-administered modes they may only hear them. Furthermore, mail mode respondents see the questions on paper, while web mode respondents see them on electronic devices with varying screen sizes. Mail mode respondents answer using a pen or pencil while web mode respondents use a mouse, keyboard or touchscreen. Such features can have an influence on the comparability and quality of the measurement instrument.

Klausch et al.'s (2013) findings suggest that comparability of measurement between modes may not be attainable when comparing self-administered and interviewer-administered modes, but that measurement between different self-administered modes is comparable. Other authors have also reported this pattern (see e.g., De Leeuw & Hox, 2011; Hox et al., 2017). Yet, there are also studies finding differences within the group of self-administered modes, more specifically between mail and web mode, on aspects such as response quality, response patterns and estimation precision (Savage & Waldman, 2008; Olsen, 2009; Kwak & Radler, 2002). These differences are, for example, hypothesised to be due to online respondents suffering more fatigue and boredom which, in turn, could be caused by visual and interactive stimuli in the online mode being more cognitively demanding (Savage & Waldman, 2008). Kwak and Radler (2002) also discuss that differences in visual display, for example, different sizes of open-answer fields, or in the relative burden caused by filter questions in mail as compared to in web surveys, could cause such mode differences. Olsen (2009) attributes these mode differences to different self-selection processes into the mode groups.

If measurement differs between modes, different ways of designing a survey question, which we refer to as different methods in the following, might thus be preferable per mode to ensure the highest measurement quality. For example, longer response scales, i.e., with more answer categories (see e.g., Alwin, 1997; Andrews, 1984; Cox III, 1980; Költringer, 1995; Saris et al., 1977), as well as fully labelled response scales (see e.g., Alwin, 2007; Alwin & Krosnick, 1991; Saris & Gallhofer, 2007) tend to lead to higher measurement quality. However, one might not expect long and fully labelled scales to lead to high measurement quality in purely oral modes where respondents are unlikely to keep all response options equally present in their memory before answering (Krosnick & Alwin, 1987). Thus, long lists of response categories are typically not read out in oral modes (Schwarz et al., 1991). For unimode surveys, method recommendations tailored to the employed mode may thus be followed. However, where comparability across modes is crucial, such as in mixed-mode surveys, the focus should be on finding those methods that lead to the highest measurement quality in all modes used.

Various question characteristics have been studied in terms of their links to measurement quality (see e.g., DeCastellarnau, 2018). While, in practice, question characteristics are often interrelated and there are no incontestable unique guidelines on what works best (DeCastellarnau, 2018; Saris & Gallhofer, 2014; Schaeffer & Dykema, 2020), this research helps questionnaire designers. It enables them to carefully consider the way they employ question characteristics in their measurement instruments, taking into account different theoretical arguments and empirical evidence. Previous research has determined the measurement quality of specific questions (see e.g., Revilla et al., 2014; Oberski et al., 2007) as well as the influence of question characteristics on measurement quality through meta-analysis (Kogovšek & Ferligoj, 2005; Saris & Gallhofer; 2014; Saris et al., 2011; Scherpenzeel & Saris, 1997). Such research has been conducted in several countries, concerning various question topics and in different modes of data collection. Still, as web surveys have existed for a relatively short time, measurement quality assessments for this mode are still rarer than for other modes (Bosch et al., 2019). Furthermore, web panels are a special context in which web surveys are administered, on which even less research exists. The specificity here comes particularly from the fact that panel conditioning, i.e., training or learning effects, can appear, which tend to lead to an increase in the reliability and stability of responses over time (Sturgis et al., 2009). Moreover, as in most countries substantive parts of the population do not use the internet (World Bank, 2020), it is crucial to also study the measurement quality of survey questions in mixed-mode settings (Callegaro et al., 2014).

This study therefore sets out to assess measurement quality in a mixed-mode panel survey, using web and mail mode, to find a measurement method that results in the highest measurement quality across both modes. We do this by conducting two Multitrait-Multimethod (MTMM) experiments, allowing us to estimate measurement quality as defined above. Furthermore, to advance research on the links between question characteristics and measurement quality, we particularly focus on the effect of two response scale characteristics, namely the length and labelling of response scales.

This paper proceeds as follows: We first present the theoretical argumentation and empirical evidence concerning using response scales of a certain length and using fully versus partially labelled response scales. On this basis, we formulate hypotheses. We then describe the data, the experimental design and the analytical strategy. Subsequently, we present the results, discuss them and draw conclusions.

Theory and Empirical Evidence: Scale Characteristics and their Impact on Measurement Quality

Length of Response Scales: Theory

Much of the literature on the relationship between response scale length and measurement quality bases its theoretical argument on the theory of information (e.g., Alwin, 2007; Alwin et al., 2018; Revilla et al., 2014). The theory of information suggests that with an increasing number of scale points, not only the direction but also the intensity or extremity of an attitude can be assessed in an increasingly detailed fashion (Garner, 1960). Therefore, longer scales should result in better measurement quality because more information can be gathered (see also Alwin, 1997; Andrews, 1984; Cox III, 1980; Költringer, 1995; Saris et al., 1977). Along the same lines, Alwin and Krosnick (1991) describe that offering too few categories would lead to a loss of information, as respondents would have to 'round' their answers.

However, there are also arguments for not including too many answer categories. Schaeffer and Presser (2003) state that the right response scale length should be a compromise between offering more potential for finer distinctions and considering respondents' limited capacities for making finer distinctions reliably and in similar ways. For example, a 100-point scale enables respondents to make finer distinctions than a five-point scale. However, it bears higher potential to induce different responses from a respondent when asked repeatedly across time, as well as to be used in different ways across respondents compared to a five-points scale. Similarly, other authors argue against the use of long response scales, referring to the suggestion of cognitive theorists that there is an upper limit to how many answer categories respondents can handle (Vall-Llosera et al., 2020). At a certain point, adding more categories results in the answer options having less rather than more meaning. Moreover, referring to motivational theories, the task of answering survey questions becomes increasingly complex the more answer categories are offered, thus too many scale points could lead to satisficing (Alwin, 1997). Especially scholars discussing very long scales have pointed out that respondents are likely to engage in rounding which can be regarded as a form of satisficing because, rather than considering all answer options, the task complexity is reduced by effectively only considering a part of the answer options (Liu & Conrad, 2016; Tourangeau et al., 2000).

Length of Response Scales: Empirical Evidence

Previous studies offer a large body of empirical findings on the optimal length of response scales. It should be noted that response scales can be classified in terms of various characteristics, such as the scales' evaluative dimension (item-specific versus agree-disagree) or the scales' polarity (unipolar versus bipolar) (DeCastellarnau, 2018). These characteristics are interrelated. For example, agree-disagree scales are always bipolar. To review the empirical literature on the impact of single response scale characteristics on measurement quality, we will draw from findings on scales that are otherwise heterogenous. For example, we will look at the effects of response scale length in both item-specific and agree-disagree scales, to gather as many findings as possible on the impact of response scale length on measurement quality. Moreover, due to the mixed-mode angle of this study, we will also consider whether results differ across modes in our review. Where operationalizations of measurement quality, reliability or validity diverge from the operationalization we use, this will be indicated in the following by specifying the exact indicator used (e.g., test-retest reliability) or by describing the operationalization.

Many scholars report an improvement of measurement quality with an increase of answer categories. For example, Alwin (1997) finds higher reliability and validity for eleven-point scales than for seven-point scales in a study employing face-to-face mode. Andrews (1984) also finds that using more categories increases measurement quality, both in terms of reliability and validity, in a study conducted in various modes, namely telephone, face-to-face and group interviews. Rodgers et al. (1992) find in a face-to-face study that both validity and reliability increase with the number of scale points. Lundmark et al. (2016) look at concurrent validity, i.e., the extent to which a variable can predict other variables it should be related to. They find this to be higher for longer scales (seven and eleven-point scales as compared to two-point scales) in a web mode survey. Furthermore, Wu and Leung (2017) use simulated survey data to compare scales of four, five, six, seven and eleven points and find the longer scales to lead to higher measurement quality, here defined as the accordance of the simulated data with the 'true scores' calculated from a known underlying distribution. Revilla and Ochoa (2015) similarly find longer scales to lead to better measurement quality, at least up to eleven points, focusing on item specific scales in a web survey. Yet, looking specifically at agree-disagree scales, Revilla et al. (2014) do not find measurement quality to improve by increasing the number of scale points beyond five. Their results are based on a face-to-face study.

Many authors find that improving measurement quality by increasing the number of answer categories only works up to a certain point beyond which no improvements are observed. Instead, quality might even decrease. This is often described as a curvilinear effect. For example, Preston and Colman's (2000) findings suggest a curvilinear effect when looking at test-retest reliability: Adding cat-

egories increases this measure of reliability between two and ten scale points, but adding further points leads to a decrease in test-retest reliability. They find a similar pattern when looking at indicators of criterion and convergent validity. Their study was conducted using self-administered paper questionnaires. Similarly, Saris and Gallhofer (2007) find that increasing the number of categories up to eleven points leads to improved measurement quality in their meta-analysis based on data from face-to-face interviews, the disk-by-mail approach[1] and the Telepanel[2]. Alwin and Krosnick (1991), using data from face-to-face interviews, find that for item specific questions, the quasi-simplex model reliability[3] increases from three to seven points and then remains constant when the scale is extended to nine points.

In contrast, other scholars find relatively short scales to be superior. McKelvie (1978) finds that test-retest reliability tends to be highest when using five-point scales in his study using self-administered paper questionnaires. Alwin (2007), looking particularly at unipolar scales and using quasi-simplex models, finds that they are most reliable at four points. He bases his findings on a mix of face-to-face and self-administered paper questionnaire surveys. Scherpenzeel and Saris (1997) stress the different effects response scale length can have on validity and reliability showing that validity is highest at four, five or seven points while reliability is highest at two to three points. They analyse data from web surveys, mail surveys and computer assisted telephone interviews (CATI). They also look at potential differences between modes but do not find any. Alwin et al. (2018) find that reliability tends to decline with an increasing number of response options, with two-point scales resulting in the highest reliability. Unipolar measures of attitudes form the exception. For this type of question, reliability increases with longer scales. Their analysis is based on General Social Survey questions conducted in face-to-face mode.

There are also studies suggesting that changing the number of response categories does not affect measurement quality. Jacoby and Matell (1971), looking at both test-retest reliability and indicators for predictive and concurrent validity, find this in their study focusing on agree-disagree scales based on self-administered paper questionnaires. McKelvie (1978) also finds indications for this, at least in terms of validity, in his study using self-administered paper questionnaires. More precisely, he does not find criterion validity, operationalized as correlating responses with

1 A floppy disk containing the survey and the programme required to open the survey was sent to respondents.

2 An early web mode approach. Respondents were provided with a computer and a modem, if necessary, so that surveys could be sent to them.

3 The quasi-simplex model is an extension of the test-retest model using at least three repeated measures of the same variable over time to estimate reliability. It allows to account for change in the measure of interest and assumes that there is no method effect (Saris & Gallhofer, 2014).

available objectively correct values, to be affected by a change in the number of answer categories.

Overall, most empirical findings seem to suggest that longer scales can indeed lead to higher measurement quality but that this only works up to a certain point from which on quality tends to remain stable. Yet, different studies find different optimal numbers of scale points, ranging from five to eleven. From this review, we cannot deduce that this should differ between web and mail mode. We therefore expect that *response scales with five to eleven points result in the highest measurement quality in both web and mail mode (H1).*

Fully Labelled Versus Partially Labelled Response Scales: Theory

More comprehensive labelling of a scale is commonly assumed to be beneficial as it clarifies the meaning of otherwise ambiguous scale points, thus reducing variability in scale point interpretation across respondents (Alwin, 2007; Eutsler & Lang, 2015; Krosnick & Berent, 1993; Krosnick & Fabrigar, 1997). Verbal labels should be a more natural form of expressing meaning compared to numbers (Krosnick & Fabrigar, 1997). Receiving the information in text form, rather than via numbers, should therefore reduce respondent burden (Krosnick & Presser, 2010).

Yet, there are arguments that suggest more extensive verbal labelling might be harmful to measurement quality. For example, Krosnick and Fabrigar (1997) mention that verbal labels could be problematic due to language ambiguity and are also more difficult to remember (see also Alwin & Krosnick, 1991). They argue that the task of answering a survey question could be less cognitively demanding for respondents if they have to read fewer labels, for example, when only end point labels are used (see also Kunz, 2015). This stands in direct contrast with the argument made above. Menold et al. (2014) reconcile these opposing assumptions stating that while full verbal labelling facilitates interpretation, it makes the mapping process more burdensome when compared to end point labelling.

Fully Labelled Versus Partially Lablled Response Scales: Empirical Evidence

The vast majority of research finds fully labelled scales to be superior to partially labelled ones of similar length. For example, Alwin (2007) finds the quasi-simplex model reliability of response scales to increase when full labels rather than just endpoint labels are used. He bases his work on a variety of face-to-face and self-administered paper questionnaire surveys (administered on site, i.e., not mail mode). Alwin and Krosnick (1991) find that using fully labelled response options is

associated with an increase in quasi-simplex model reliability in a study based on face-to-face and telephone surveys. Similarly, Saris and Gallhofer (2007), in their study based on data from face-to-face interviews, the disk-by-mail approach and the Telepanel find that the use of verbal labels increases the reliability of questions.

There are, however, also some findings that point in the opposite direction. Andrews (1984) concludes from his analyses of data collected in telephone and face-to-face individual and group interviews that measurement quality decreases where fully labelled answer categories are used. Similarly, Rodgers et al. (1992) find full labelling to lead to more random measurement error, i.e., lower reliability, in a face-to-face survey.

To sum up, most empirical assessments of the issue find that fully labelled scales lead to higher measurement quality. This was found to be the case across various modes. We therefore expect that *fully labelled response scales lead to higher measurement quality in both web and mail mode* (H2).

Comparing the Effects of Scale Length and Full Labelling

So far, we have focused on the impact of the length of response scales and the labelling of response scales separately. However, for the sake of deriving practical recommendations for questionnaire designers, we would also like to assess whether it is more beneficial for measurement quality to have a long or a fully labelled response scale. We could only find one study that compared the effect of these two characteristics on measurement quality based on a meta-analysis. Andrews (1984) shows that the number of scale categories explains a larger share of the variance in validity and reliability than the labelling of the scale. Therefore, we expect that *the benefit of employing longer response scales will outweigh the benefit of employing fully labelled response scales* (H3).

Data and Method

Sample

We conduct the experiments in the GESIS Panel, a probability-based mixed-mode panel in which about 75% of the respondents answer in web mode and 25% in mail mode. The GESIS Panel was founded in 2013 and contains about 5000 panelists. To account for attrition, the sample was refreshed in 2016 and 2018. Every two months, panelists are invited to participate in a survey lasting approximately 20 minutes. They receive a five-euro prepaid incentive with each survey invitation (GESIS, 2020; Minderop et al., 2019; Bosnjak et al., 2017). Upon a face-to-face recruitment interview, those respondents who indicated that they use the internet regularly were

Table 1 Characteristics of sample before listwise deletion for both web and mail mode (unweighted)

	Web mode			Mail mode		
	Mean	SD	Valid n	Mean	SD	Valid n
Age	47.08	14.36	2,779	57.19	12.69	1,028
Female	49.14%	.50	2,784	54.07%	49.86	1,032
University education	34.18%	47.44	2.762	13.76%	34.46	1,025
Total			2,784			1,032

offered to participate in web mode. Interviewers were requested to present online participation as an attractive option and to persuade respondents to participate in web mode. However, internet users were also free to opt for mail mode. Those respondents who did not use the internet were only presented the option to participate in mail mode (Bosnjak et al., 2017). The Multitrait-Multimethod experiments were implemented in the 'gb' wave fielded in April and May 2019 (GESIS, 2020).

Sociodemographic characteristics of both web mode and mail mode respondents in the sample before listwise deletion[4] of respondents with missing values on the experimental variables are presented in Table 1. As can be expected, respondents who self-selected into the web mode differ significantly from those who self-selected into the mail mode. Mail respondents are on average about ten years older than web respondents ($p<.001$). Furthermore, the proportion of female respondents is about five percentage points higher among mail respondents than among web respondents ($p<.05$). Women are thus overrepresented in mail mode. The proportion of respondents who have obtained a university degree is substantially higher among web mode respondents (34%) than among mail mode respondents (14%) ($p<.001$). After listwise deletion of cases with missing values, the total valid sample size for experiment 1 (environmental attitudes) is n=3,632 and n=3,589 for experiment 2 (supernatural beliefs). We also conduct analyses of variance to check if sociodemographic characteristics differ significantly across the experimental groups. The results show that differences approach significance ($p=.0596$) only for 'university education'. Concretely, the proportion of respondents who indicated that a university degree is their highest achieved level of education is about four percentage points lower for group two (26.34%) than for groups one and three (30.29% and 29.34%, respectively). As this difference is substantively small, we see no rea-

4 We ran a robustness analysis using pairwise deletion instead. The resulting estimates are extremely similar to those found using listwise deletion. The results would not lead to an alteration of any substantive findings.

son to be concerned about the success of respondents' random assignment into experimental groups.

The True Score MTMM Model

The MTMM experimental design used here is based on the True Score MTMM (TS-MTMM) model proposed by Saris and Andrews (1991) to estimate the reliability, validity, and quality of the survey questions. According to Saris and Andrews (1991), measurement quality is defined as the product of validity and reliability. Validity is defined as the strength of the relationship between a latent variable of interest and the 'true score' and reliability as the strength of the relationship between the 'true score' and the observed variable.

The following system of equations describes the TS-MTMM model:

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \tag{1}$$

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \tag{2}$$

with F_i being the i^{th} trait or factor, M_j being the j^{th} method, Y_{ij} being the observed answer for the i^{th} trait and the j^{th} method, T_{ij} being the true score factor or systematic component of the response, r_{ij} being the reliability coefficient (when standardized), v_{ij} being the validity coefficient (when standardized), and e_{ij} being the random error associated with Y_{ij}.

Equation (1) defines the observed variables as the sum of the associated systematic component and random errors. Equation (2) defines the systematic components themselves as the sum of the trait component and the effect of the method employed to assess the trait. The total measurement quality can be obtained by taking the product of the reliability and validity, being the reliability coefficient and the validity coefficient squared: An illustration of the path diagram of the True Score MTMM model for three traits, each measured with three methods is presented in Figure 1.

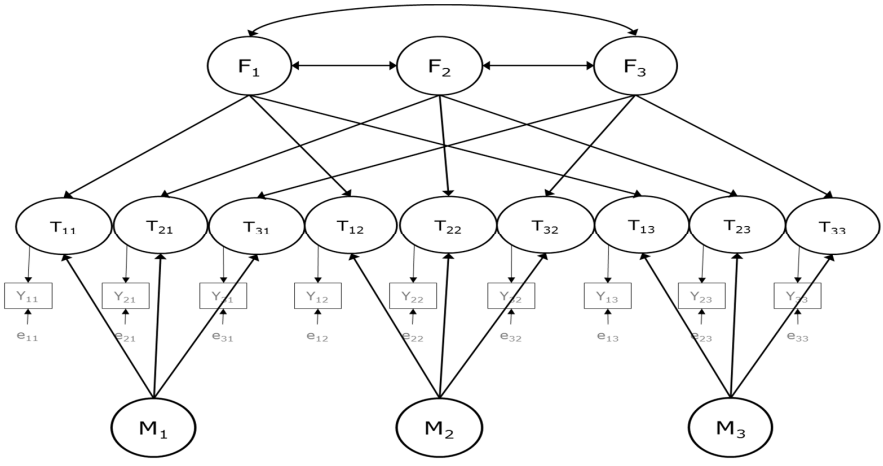


Figure 1 Path diagram of the True Score MTMM model for three traits and three methods

It shows that each trait (F_i) is measured three times with different methods (M_j). This results in nine true scores (T_{ij}) which are measured by the nine survey questions that are evaluated in each experiment. The observed responses to these nine questions are denoted as Y_{ij}. By measuring three correlated traits with three methods, we can thus estimate the measurement quality of all employed survey questions estimating the TS-MTMM model using Structural Equation Modelling (SEM) (see also section on analytical strategy).

As Figure 1 shows, we assume the traits (F_i) to be correlated, the method factors (M_j) to be uncorrelated, and the method factors to be uncorrelated with the trait factors. We also assume that the impact of the method factor on the traits measured with a common scale is the same and that the random errors (e_{ij}) are uncorrelated with each other and with the true scores (T_{ij}), the trait factors (F_i) and the method factors (M_j).

The Assessed Traits

The traits for experiment 1 are three questions on environmental attitudes based on previous questions asked on the GESIS Panel (GESIS, 2020). The traits for experiment 2 are three questions in supernatural beliefs based on questions from the ALLBUS 2012 (GESIS, 2016). Table 2 shows English translations of the questions. The German questions can be found in Appendix B.

Table 2 Traits

Experiment 1: Environmental attitudes	
Trait 1	Can you identify with environmentalists?
Trait 2	Should we all be willing to restrict our current living standard for the benefit of the environment?
Trait 3	Do you believe that some problems of our times would be solved if we went back to a more rural and natural lifestyle?
Experiment 2: Supernatural beliefs	
	How much do you believe in the following?
Trait 1	…in life after death
Trait 2	…in heaven
Trait 3	…in miracles

The Assessed Methods

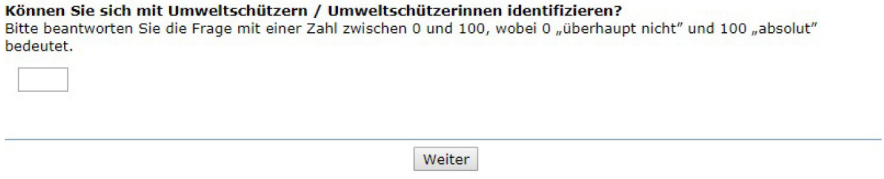
To test our hypotheses, we focus on varying the length of the response scales and the extent of labelling answer categories. However, to be able to identify the MTMM model in the analysis, it is helpful to vary further question characteristics. In the three assessed methods, we vary the following characteristics (see also Table 3): (1) the length of the response scale; (2) the verbal labelling of the response scale (fully versus partially labelled); (3) whether a continuous or discrete scale is used; (4) whether the scale is presented in a horizontal format or as a numerical open-ended scale; (5) whether a definition of the scale is present in the request or not. Figures 2 and 3 display how the methods for the first trait of the first experiment appear in the GESIS Panel web and mail questionnaire, respectively. In Appendix A, we present an exemplary smartphone screenshot, showing that the horizontal response scales were also displayed horizontally on small screen mobile devices.

Table 3 Variations of question characteristics across methods

Variation	Method 1	Method 2	Method 3
1	7-points	101-points	11-points
2	Fully labelled	Partially labelled	Partially labelled
3	Discrete	Continuous	Discrete
4	No definition of scale	Definition of scale	Definition of scale
5	Horizontal	Numerical open-ended scale	Horizontal

Method 1:

Method 2:



Method 3:

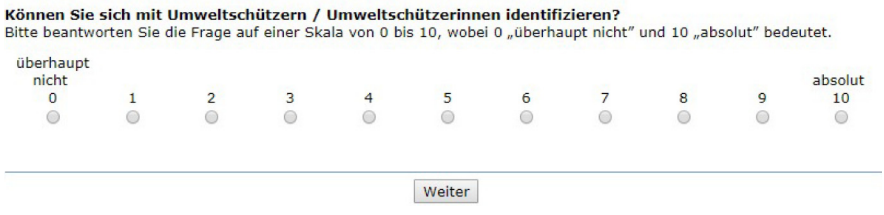


Figure 2 Screenshots of the GESIS Panel web questionnaire: Trait 1 of experiment 1 asked with methods 1, 2 and 3

Method 1:

(50) Können Sie sich mit Umweltschützern / Umweltschützerinnen identifizieren?

überhaupt nicht	ein wenig	etwas	mäßig	erheblich	sehr	absolut
o	o	o	o	o	o	o

Method 2:

(50) Können Sie sich mit Umweltschützern / Umweltschützerinnen identifizieren?

Bitte beantworten Sie die Frage mit einer Zahl zwischen 0 und 100, wobei 0 „überhaupt nicht" und 100 „absolut" bedeutet.

___

Method 3:

(50) Können Sie sich mit Umweltschützern / Umweltschützerinnen identifizieren?

Bitte beantworten Sie die Frage auf einer Skala von 0 bis 10, wobei 0 „überhaupt nicht" und 10 „absolut" bedeutet.

überhaupt nicht										absolut
0	1	2	3	4	5	6	7	8	9	10
o	o	o	o	o	o	o	o	o	o	o

Figure 3 Depiction of the GESIS Panel mail questionnaire: Trait 1 of experiment 1 asked with methods 1, 2 and 3

Experimental Design

For the experiments, respondents are randomly assigned to three groups of approximately equal size. In three-group Split Ballot MTMM experiments, each group receives three questions asking for the three traits using one method at time 1 (towards the middle of the questionnaire), and each group receives the same three questions again but with another method at time 2 (at the end of the questionnaire). Other questions are asked in between the two instances to reduce memory effects (Schwarz et al., 2020; Van Meurs & Saris, 1990). By implementing two methods in each group but varying which methods these are across groups, all combinations of methods are covered. Also, respondents do not have to handle the burden and potential fatigue that would result from asking them the same questions three times. As we run two MTMM experiments in one survey, we vary which groups are asked with which methods across the two experiments (see Table 4) to avoid repetitions of the same methods within a group as much as possible.

Table 4 Three-group Split-Ballot MTMM Design for both experiments

	Time 1		Time 2	
	Experiment 1	Experiment 2	Experiment 1	Experiment 2
Group 1	M1	M2	M2	M3
Group 2	M2	M3	M3	M1
Group 3	M3	M1	M1	M2

Analytical Strategy: Model Estimation and Testing

For model estimation we use Maximum Likelihood in LISREL 8.72 (Jöreskog & Sörbom, 1996). The base model in LISREL notation can be found in Appendix C. We run three separate analyses: (1) for the entire sample, (2) only for web mode respondents and (3) only for mail mode respondents. For testing, we evaluate the local model fit with the software JRule (Van der Veld et al., 2008). Parameter misspecifications indicated by JRule are used to improve the model. Such improvements can consist in allowing unequal effects of one method on the different traits, freeing error variances because of timing effects, adding a correlation between two methods, or allowing correlations between errors due to expected memory effects. As we expect the same models to hold in the analysis of the entire sample as well as presenting reliability and validity separately, we aim to implement the same adjustments to the model across these analyses. However, this is not always possible (i.e., it can result in improper solutions or poor model fit). The final model adjustments for all analyses are shown in Appendix D, as are the global model fit indices and indications of remaining local misspecifications as shown by JRule.

Results

In Table 5, we present the average measurement quality across traits by method, experiment and mode. The detailed results, i.e., per trait as well as presenting reliability and validity separately, are shown in Appendix E. We consider a quality estimate above or equal to .9 to indicate excellent measurement quality, a quality estimate between .9 and .8 good quality and a quality estimate between .8 and .7 acceptable quality. A quality estimate between .7 and .6 is seen as questionable, and quality estimates below .6 are interpreted as poor measurement quality.

Table 5 Average quality across all traits by method, mode of data collection and experiment

	Experiment: Environmental Attitudes			Experiment: Supernatural Beliefs			Both experiments		
	Both modes	Web	Mail	Both modes	Web	Mail	Both modes	Web	Mail
M1	0.69	0.73	0.59	0.90	0.89	0.89	0.80	0.81	0.74
M2	0.72	0.77	0.66	0.89	0.90	0.86	0.80	0.83	0.76
M3	0.83	0.83	0.79	0.99	1.00	0.97	0.91	0.91	0.88

Note: M= Method; M1= 7-point fully labelled horizontal, no scale definition; M2= 101-point numerical open-ended scale, scale definition present; M3= 11-point partially labelled horizontal, scale definition present.

Looking at all average quality estimates across methods and experiments, we find that measurement quality tends to be especially high for the experiment on supernatural beliefs, with all estimates indicating excellent or good quality. Findings are more mixed for the experiment on environmental attitudes, with quality estimates ranging from good to questionable and even to poor in one instance (for method 1 in mail mode).

When we look at the average quality for each method in both experiments, we find that, overall, method 3 (eleven points, only end points labelled) obtains the highest quality (between .79 and 1), independently of the mode of data collection or the topic of the experiment. Thus, our hypothesis that the benefits of using long response scales outweigh the benefits of using fully labelled response scales (H3) cannot be rejected. Comparing the performance of method 1 (seven-point fully labelled) and method 2 (numerical open-ended scale ranging from zero to 100) in all modes and experiments shows that they perform similarly. An exception can be observed in mail mode in experiment 1, where method 1 performs substantially worse than method 2. The similar performance of methods 1 and 2 is not in line with our expectation formulated in H1 that scales between five and eleven points result in the highest measurement quality. Instead, our results show that the 101-point scale tends to result in the same measurement quality as the seven-point scale. For the experiment on environmental attitudes, it appears that the longer scale even outperforms the seven-point scale, at least for mail mode. Moreover, the observation that method 1 results in the lowest measurement quality in most instances and that it is consistently outperformed by the partially labelled scale (method 3) means we can reject H2 that fully labelled response scales lead to higher measurement quality.

Furthermore, the observation that method 3 performs best across all modes and that there are few differences in the performance of methods 1 and 2 in the different modes also means that it is indeed possible to find one method that performs best across both modes in this case[5].

Discussion and Conclusion

In this paper we set out to assess which response scale results in the highest measurement quality across two modes of data collection, self-completion in a web survey (web mode) and on a paper questionnaire (mail mode). Given the differing internet penetration and internet literacy across and even within countries, such mixed-mode designs are a valuable option to increase survey participation while saving costs. Based on the state-of-the-art in the field, we formulated hypotheses regarding the impact of length and labelling of response scales on measurement quality.

In line with the literature, we find that the eleven-point partially labelled scale (method 3) consistently produces the highest measurement quality across modes for both experiments (Preston & Colman, 2000; Rodgers et al., 1992; Saris & Gallhofer, 2007). Contrary to previous results, we find that a numerical open-ended scale, i.e., a scale requiring respondents to indicate the answer using a number, here between zero and 100, and a seven-point fully labelled response scale tend to result in the same measurement quality. Previous literature has found fully labelled scales to lead to higher measurement quality across various modes including self-completion on the web (Saris & Gallhofer, 2007) and on paper questionnaires (Alwin, 2007).

Moreover, differences in measurement quality across modes have been reported (Sánchez Tomé, 2018; Tourangeau, 2017; Dillman et al., 2014). However, our study suggests that there are no systematic differences across modes concerning the effect of response scale length and labelling on measurement quality.

Furthermore, we find that using longer response scales seems to give more of a boost to measurement quality than using fully labelled scales (H3). The partially labelled eleven-point scale (method 3) outperforms the fully labelled seven-point scale (method 1) consistently, and the numerical open-ended scale (method 2) outperforms the fully labelled seven-point scale (method 1) for one mode in one experiment. However, longer scales are not generally better. Our study shows that

5 We also ran a robustness analysis on only respondents using smartphones (valid n for the experiment on environmental attitudes is 512 and for the experiment on supernatural beliefs is 516). The resulting estimates are extremely similar to those found for web mode overall. Analysing smartphone respondents separately leads to the same substantive findings.

increasing the number of scale points from seven to eleven yields higher measurement quality but increasing it from eleven to 101 points leads to inferior measurement quality.

On the basis of these findings, we can recommend using an eleven-point partially labelled scale (method 3) when measuring attitudes or beliefs in mixed-mode surveys combining web and mail mode. Furthermore, we recommend prioritizing the use of longer response scales (up to eleven points) over the use of seven-point fully labelled scales.

One limitation of our study results from the suboptimal formulation of the questions of experiment 1. Question formulations here did not indicate that respondents would be able to give a nuanced answer but read as yes/no questions. This might partly explain why lower measurement quality is obtained by the questions of experiment 1 compared to those of experiment 2.

Another limitation is that, given the design of our experiments, we cannot draw conclusions beyond the particular combinations of characteristics present in the tested response scales. To estimate an MTMM model, several scale characteristics should be varied across methods. Therefore, we could not assess the isolated effect of one scale characteristic. To do so, more experiments and a meta-analysis are needed (Kogovšek & Ferligoj, 2005; Saris & Gallhofer, 2014; Saris et al., 2011; Scherpenzeel & Saris, 1997). However, in terms of practical implications it is not always necessary to unconfound the impact of different question characteristics. In survey practice, specific question characteristics tend to occur together (e.g., eleven-point scales are usually only partially labelled), while the combination of other question characteristics is less practically feasible, less common, and therefore less relevant to study (e.g., eleven-point scales are rarely fully labelled). These "structural dependencies among sets of characteristics" are also pointed out by Schaeffer and Dykema (2020, p.10.6), reminding us that decisions in the design of survey questions depend on what combinations of characteristics can or cannot occur together. The results of MTMM experiments showing which measurement scales lead to which measurement quality thus remain a vital basis for questionnaire design.

Further research is needed to investigate the questions left open by this study: Does full labelling only lead to higher measurement quality in shorter scales? Are eleven points really the optimal length, or may slightly shorter scales (for example: nine points) be preferable? If the seven-point scale had been only partially labelled, would it still be outperformed by the partially labelled eleven-point scale? And would any of these adjustments have resulted in differences across modes? In short, a variety of feasible scales remain to be tested on mixed-mode panels such as the GESIS Panel and mode differences should always be taken into account.

Bibliography

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods and Research*, *25*(3), 318-340. doi:10.1177/0049124197025003003

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods and Research*, *20*(1), 139–181. doi:10.1177/0049124191020001005

Alwin, D. F., Baumgartner, E. M., & Beattie, B. A. (2018). Number of response categories and reliability in attitude measurement. *Journal of Survey Statistics and Methodology*, *6*(2), 212-239. doi:10.1093/jssam/smx025

Alwin, D.F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken: Wiley.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, *48*(2), 409-442. doi:10.1086/268840

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A. S., Das, M., Douhou, S., & Krieger, U. (2016). A Comparison of four probability-based online and mixed-mode Panels in Europe. *Social Science Computer Review, 34*(1), 8-25. doi:10.1177/0894439315574825

Bosch, O. J., Revilla, M., DeCastellarnau, A., & Weber, W. (2019). Measurement reliability, validity, and quality of slider versus radio button scales in an online probability-based panel in Norway. *Social Science Computer Review, 37*(1), 119-132. doi:10.1177/0894439317750089

Bosnjak, M., Dannwolf, T., Enderle, T., Schauer, I., Struminskaya, B., Tanner, A., & Weyandt K. W. (2017). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review, 36*(1). doi:10.1177/0894439317697949

Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (2014). Online panel research: History, concepts, applications and a look at the future. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick & P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 1-22). Hoboken: Wiley.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81-105.

Cox III, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*(4), 407-422. doi:10.1177/002224378001700401

De Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as part of a mixed-mode design. In M. Das, P. Ester & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (pp. 45-76). New York: Routledge.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, *52*(4), 1523-1559. doi:10.1007/s11135-017-0533-4

Dillman, D. A., Smyth, J. D. & Christian, L. M. (2014) Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. Hoboken: Wiley.

Eifler, S., & Faulbaum, F. (2017). Vorwort. In S. Eifler & F. Faulbaum (Eds.), *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 7-8). Wiesbaden: Springer VS.

ESOMAR & WAPOR (2014). ESOMAR/WAPOR Guideline on Opinion Polls and Published Surveys (World Research Codes and Guidelines). Retrieved from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-WAPOR-Guideline-on-Opinion-Polls-and-Published-Surveys-August-2014.pdf

Eutsler, J., & Lang, B. (2015). Rating scales in accounting research: The impact of scale points and labels. *Behavioral Research in Accounting, 27*(2), 35-51. doi:10.2308/bria-51219

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review, 67*(6), 343-352. doi:10.1037/h0043047

GESIS (2016). *ALLBUS 1980-2014 – Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. ZA4582 Datenfile Version 1.0.0*, [Data file] Gesis – Leibniz Institute for the Social Sciences. doi:10.4232/1.12439

GESIS (2020): *GESIS Panel - Standard Edition. GESIS Data Archive, Cologne. ZA5665 Datenfile Version 33.0.0*, [Data file] Gesis – Leibniz Institute for the Social Sciences. doi:10.4232/1.13377

Grewenig, E., Lergetporer, P., Simon, L., Werner, K., & Woessmann, L. (2018). *Can online surveys represent the entire population?* (CESifo Working Paper No. 7222). Retrieved from https://ssrn.com/abstract=3275396

Hox, J. de Leeuw, E. & Klausch, T. (2017) Mixed mode research: Issues in design and analysis. In Biemer, P., de Leeuw, E. Eckman, S., Edwards, B., Kreuter, F.,Lyberg, L. E., Tucker, N. C. & West, B.T. (Eds.), *Total Survey Error in Practice* (pp. 511-530). Hoboken: Wiley.

Jacoby, J., & Matell, M. S. (1971). Three-point Likert Scales are good enough. *Journal of Marketing Research*, *8*(4), 495–500. doi:10.2307/3150242

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Uppsala, Sweden: Scientific Software International.

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research, 42*(3), 227-263. doi:10.1177/0049124113500480

Kogovšek, T., & Ferligoj, A. (2005). The quality of measurement of personal support subnetworks. *Quality and Quantity, 38*(5), 517-532. doi:10.1007/s11135-005-2178-y

Költringer, R. (1995). Measurement quality in Austrian personal interview surveys. In W. E. Saris & A. Münnich (Eds.), *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments* (pp. 207–224). Budapest: Eötvös University Press.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201-219.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identifications and policy preferences: the impact of survey question format. *American Journal of Political Science*, *37*(3), 941–964. doi:10.2307/2111580

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141-164). Hoboken: Wiley.

Krosnick, J.A., & Presser, S. (2010). Question and questionnaire design. In P.V. Marsden, & J.D. Write (Eds.), *Handbook of Survey Research* (pp. 263–313). Bingley: Emerald.

Kunz, T. (2015). *Rating scales in web surveys. A test of new drag-and-drop rating procedures* (Doctoral dissertation). Retrieved from https://tuprints.ulb.tu-darmstadt.de/5151/7/Kunz_2015_Rating_scales_in_web_surveys.pdf

Kwak, N., & Radler, B. (2002). A comparison between mail and web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics, 18*(2), 257.

Liu, M., & Conrad, F. G. (2016). An experiment testing six formats of 101-point rating scales. *Computers in Human Behavior, 55*, 364-371. doi:10.1016/j.chb.2015.09.036

Lundmark, S., Gilljam, M., & Dahlberg, S. (2016). Measuring generalized trust. An examination of question wording and the number of scale points. *Public Opinion Quarterly*, *80*(1), 26–43. doi:10.1093/poq/nfv042

McKelvie, S. J. (1978). Graphic rating scales - How many categories? *British Journal of Psychology, 69*(2), 185-202. doi:10.1111/j.2044-8295.1978.tb01647.x

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods, 26*(1), 21-39. doi:10.1177/1525822X13508270

Minderop, I., Bretschi, D., Kolb, J., & Heycke, T. (2019). *GESIS Panel Wave Report Wave gb* (April/May 2019). Gesis – Leibniz Institute for the Social Sciences.

Oberski, D., Saris, W. E., & Hagenaars, J. (2007). Why are there differences in measurement quality across countries. In G. Loosveldt, M. Swyngedouw & B. Cambré (Eds.), *Measuring Meaningful Data in Social Research* (pp. 281-300). Leuven: Acco.

Olsen, S. B. (2009). Choosing between internet and mail survey modes for choice experiment surveys considering non-market goods. *Environmental and Resource Economics, 44*(4), 591-610. doi: 10.1007/s10640-009-9303-7

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1–15. doi:10.1016/S0001-6918(99)00050-5

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods and Research, 43*(1), 73-97. doi:10.1177/0049124113509605

Revilla, M., & Ochoa, C. (2015). Quality of different scales in an online survey in Mexico and Colombia. *Journal of Politics in Latin America*, *7*(3), 157–177. doi:10.1177/1866802X1500700305

Rodgers, W. L., Andrews, F. M., & Herzog, A. R. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, *8*(3), 251–275.

Sánchez Tomé, R. (2018). The impact of mode of data collection on measures of subjective wellbeing (Doctoral dissertation). Retrieved from https://serval.unil.ch/resource/serval:BIB_F89D8660FBE7.P001/REF

Saris W. E., Bruinsma, C., Schoots, W., & Vermeulen, C. (1977). The use of magnitude estimation in large scale survey research. *Mens en Maatschappij, 52* (4), 369–395.

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In Biemer, P., Groves, R. E., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.), *Measurement Error in Surveys* (pp. 575-597). Hoboken: Wiley.

Saris, W. E., & Gallhofer, I. (2007) Design, Evaluation, and Analysis of Questionnaires for Survey Research. Hoboken: Wiley.

Saris, W. E., & Gallhofer, I. (2014) Design, Evaluation, and Analysis of Questionnaires for Survey Research. Hoboken: Wiley.

Saris, W., Oberski, D., Revilla, M., Zavala-Rojas, D., Lilleoja, L., Gallhofer, I., & Gruner, T. (2011). *The development of the Pprogram SQP 2.0 for the prediction of the quality of survey questions* (RECSM Working Paper No. 24). Retrieved from https://www.upf.edu/documents/3966940/3986764/RECSM_wp024.pdf

Savage, S. J., & Waldman, D. M. (2008). Learning and fatigue during choice experiments: a comparison of online and mail survey modes. *Journal of Applied Econometrics, 23*(3), 351-371. doi:10.1002/jae.984

Schaeffer, N. C., & Dykema, J. (2020). Advances in the science of asking questions. *Annual Review of Sociology, 46,* 37-60. doi:10.1146/annurev-soc-121919-054544

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65-88. doi:10.1146/annurev.soc.29.110702.110112

Scherpenzeel, A.C., & Saris, W.E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods Research, 25*(3), 341–383. doi:10.1177/0049124197025003004

Schwarz, H., Revilla, M., & Weber, W. (2020). Memory effects in repeated survey questions: Reviving the empirical investigation of the independent measurements assumption. *Survey Research Methods, 14*(3). doi:10.18148/srm/2020.v14i3.7579

Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*(3), 193-212.

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. In Lynn, P. (Ed.), *Methodology of longitudinal surveys* (pp. 113-126). Hoboken: Wiley.

Tourangeau, R. (2017). Mixing modes. In Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C. & West, B.T. (Eds.), *Total Survey Error in Practice* (pp. 115-132). Hoboken: Wiley.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press.

Vall-Llosera, L., Linares-Mustarós, S., Bikfalvi, A., & Coenders, G. (2020). A comparative assessment of graphic and 0–10 rating scales used to measure entrepreneurial competences. *Axioms, 9*(21). doi:10.3390/axioms9010021

Van der Veld, W. M., Saris, W. E., & Satorra, A. (2008). Judgement rule aid for structural equation models version 3.0.4 beta [computer software].

Van Meurs, A., & Saris, W. E. (1990). Memory effects in MTMM studies. In A. Van Meurs. & W. E. Saris (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies* (pp. 134–146). Amsterdam: North Holland.

World Bank (2020). Individuals using the internet (% of population). Retrieved October 15, 2020, from https://data.worldbank.org/indicator/IT.NET.USER.ZS

Wu, H., & Leung, S. O. (2017). Can Likert Scales be treated as interval scales? A simulation study. *Journal of Social Service Research*, *43*(4), 527–532. doi:10.1080/01488376.2017.1329775

Yang, W., Moon, H. J., & Jeon, J. Y. (2019). Comparison of response scales as measures of indoor environmental perception in combined thermal and acoustic conditions. *Sustainability*, *11*(14), 3975. doi:10.3390/su11143975

Appendices

Appendix A:
Example smartphone screenshot

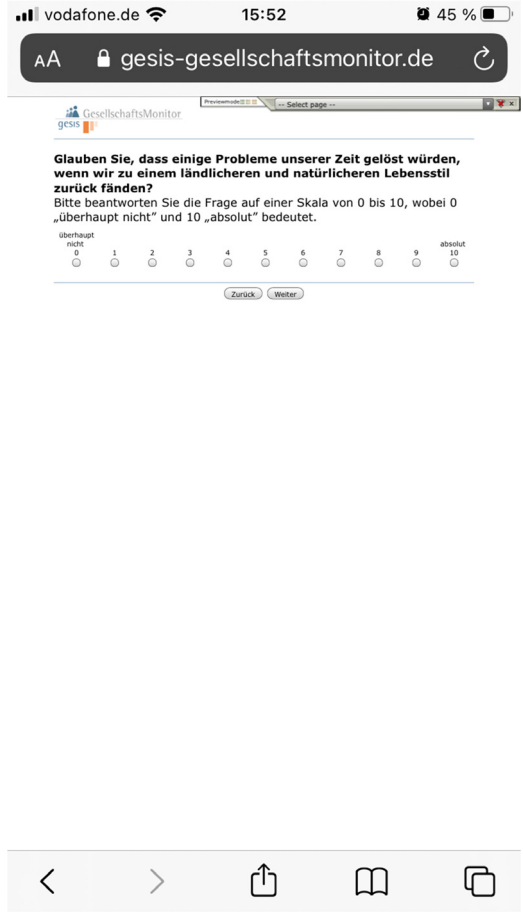


Figure 4 Depiction of the GESIS Panel web questionnaire on a smartphone: Showing the horizontal scale of method 3 (example: Experiment 1, trait 3)

Appendix B: Question formulations (original German versions)

Traits Experiment 1: Environmental attitudes

Trait 1	Können Sie sich mit Umweltschützern identifizieren?
Trait 2	Sollten wir alle bereit sein, unseren derzeitigen Lebensstandard zugunsten der Umwelt einzuschränken?
Trait 3	Glauben Sie, dass einige Probleme unserer Zeit gelöst würden, wenn wir zu einem ländlicheren und natürlicheren Lebensstil zurück fänden?

Traits Experiment 2: Supernatural beliefs

	Wie sehr glauben Sie an Folgendes?
Trait 1	An ein Leben nach dem Tod
Trait 2	An den Himmel
Trait 3	An Wunder

Appendix C:
Lisrel Input Base Model

```
! group 1
Data ng=3  ni=9 no=1368 ma=cm
km file=sb-group-1-corr.corr
mean file=sb-group-1-mean.mean
sd  file=sb-group-1-sd.sd
model ny=9 ne=9 nk=6  ly=fu,fi te=sy,fi ps=sy,fi be=fu,fi ga=fu,fi ph=sy,fi

! set lambdas of observed traits to 1, of not observed to 0
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6
value 0 ly 7 7 ly 8 8 ly 9 9

! free error variances of all observed traits, set error variance of not observed to 1
fr te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
value 1 te 7 7 te 8 8 te 9 9

! free trait gammas
fr ga 1 1 ga 2 2 ga 3 3 ga 4 1 ga 5 2 ga 6 3 ga 7 1 ga 8 2 ga 9 3

! set method gammas to 1
value 1  ga 2 4 ga 5 5 ga 8 6 ga 1 4  ga 4 5 ga 7 6
value 1 ga 3 4 ga 6 5 ga 9 6

! set trait variances to 1
value 1 ph 1 1 ph 2 2 ph 3 3

! free correlations among traits
fr ph 2 1 ph 3 1 ph 3 2

! free method variances
fr ph 4 4 ph 5 5 ph 6 6

pd
out mi iter= 5000 adm=off sc ec

! group 2
Data ni=9 no=1357  ma=cm
km file=sb-group-2-corr.corr
```

```
mean file=sb-group-2-mean.mean
sd  file=sb-group-2-sd.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in

! set lambdas of observed traits to 1, of not observed to 0
va 1 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9
value 0 ly 1 1 ly 2 2 ly 3 3

! free error variances of all observed traits, set error variance of not observed to 1
fr te 4 4 te 5 5 te 6 6 te 7 7 te 8 8 te 9 9
va 1 te 1 1 te 2 2 te 3 3

equal te 1 4 4 te 4 4
equal te 1 5 5 te 5 5
equal te 1 6 6 te 6 6

pd
out mi iter= 5000 adm=off sc ec

! group 3
Data ni=9 no=923 ma=cm
km file=sb-group-3-corr.corr
mean file=sb-group-3-mean.mean
sd  file=sb-group-3-sd.sd
model ny=9 ne=9 nk=6 ly=fu,fi te=sy,fi ps=in be=in ga=in ph=in

fr te 1 1 te 2 2 te 3 3 te 7 7 te 8 8 te 9 9
va 1 ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9 te 4 4 te 5 5
va 1 te 6 6

value 0 ly 4 4 ly 5 5 ly 6 6

equal te 1 1 1 te 1 1
equal te 1 2 2 te 2 2
equal te 1 3 3 te 3 3
equal te 2 7 7 te 7 7
equal te 2 8 8 te 8 8
equal te 2 9 9 te 9 9

pd
out mi iter= 5000 adm=off sc ec
```

Appendix D:
Final Lisrel model adjustments, fit and JRule evaluation

Experiment	Mode	Model adjustments (in LISREL notation)	df	χ2	p-value	RMSEA	CFI	JRule
Environmental attitudes	Both	FR GA14 GA76 PH44(G3)	108	70.43	0,998	0,00	1,00	None
	Web	FR GA14 GA76 PH44(G3)	108	79.56	0,982	0,00	1,00	2
	Mail	FR GA14 GA45 GA76	108	78.95	0,984	0,00	1,00	6
Super natural beliefs	Both	FR TE66(G2)	110	297.19	0,000	0,04	0,99	4
	Web	VA 0 TE99(G3)*	112	231.61	0,000	0,04	0,99	2
	Mail	FR GA34	110	108.57	0,521	0,00	1,00	None

**Note*: When looking for a suitable model to analyse the answers of online respondents in Experiment 2, the best solutions found still resulted in a small negative error variance of the observed variable measuring trait 3 with method 3 (te 9 9), equal to -.01. However, given the fact that fixing this parameter to zero neither substantially affects the resulting estimates nor the fit of the model, we decided to accept the model with this parameter fixed to zero as our final solution in this case.

Appendix E:
Reliability, validity, and quality estimates for the different traits and methods for both experiments by mode

	Reliability				Validity				Quality			
	T1	T2	T3	Avg	T1	T2	T3	Avg	T1	T2	T3	Avg
Experiment: Environmental Attitudes												
Both modes												
M1 (Time 1)	.76	.76	.81	.77	.96	.90	.92	.93	.73	.68	.75	.72
M1 (Time 2)	.77	.79	.83	.80	.92	.76	.83	.84	.71	.60	.69	.67
M2	.81	.83	.86	.83	.86	.85	.86	.86	.70	.70	.75	.72
M3	.85	.83	.85	.84	1.00	.98	.98	.99	.85	.81	.83	.83
Web												
M1 (Time 1)	.77	.79	.83	.80	.98	.90	.94	.94	.76	.71	.78	.75
M1 (Time 2)	.77	.81	.85	.81	.94	.79	.86	.87	.73	.64	.73	.70
M2	.86	.85	.88	.86	.90	.86	.90	.89	.78	.73	.80	.77
M3	.86	.83	.85	.85	.98	.98	.98	.98	.85	.81	.83	.83
Mail												
M1	.77	.72	.77	.76	.86	.71	.77	.78	.67	.51	.60	.59
M2	.76	.81	.86	.81	.77	.83	.85	.82	.59	.67	.73	.66
M3	.86	.83	.83	.84	1.00	.90	.90	.94	.86	.75	.75	.79
Experiment: Supernatural Beliefs												
Both modes												
M1	.96	.96	.94	.95	.94	.94	.94	.94	.90	.90	.89	.90
M2 (Time 1)	.96	.94	.94	.95	.96	.94	.94	.95	.92	.89	.89	.90
M2 (Time 2)	.96	.94	.88	.93	.96	.94	.94	.95	.92	.89	.83	.88
M3	.98	1.00	.98	.99	1.00	1.00	1.00	1.00	.98	1.00	.98	.99
Web												
M1	.96	.96	.92	.95	.94	.94	.94	.94	.90	.90	.87	.89
M2	.96	.94	.94	.95	.96	.94	.94	.95	.92	.89	.89	.90
M3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mail												
M1	.98	.96	.90	.95	.94	.92	.94	.93	.92	.89	.85	.89
M2	.94	.94	.88	.92	.94	.92	.92	.93	.89	.87	.81	.86
M3	.98	1.00	1.00	.99	.98	.98	.98	.98	.96	.98	.98	.97

Note: M=Method, T=Trait, M1=7-point fully labelled horizontal, no scale definition; M2=101-point numerical open-ended scale, scale definition present; M3=11-point partially labelled horizontal, scale definition present; Avg=Average.

methods, data, analyses | 2021, Vol. 15(2), pp. 191-214 DOI: 10.12758/mda.2021.01

The Measurement Invariance of Customer Loyalty and Customer Experience across Firms, Industries, and Countries

Timothy B. Gravelle
SurveyMonkey

Abstract

Research on cross-national (and cross-group) measurement invariance is now well developed in the social and behavioural sciences, but this research has yet to engage research practitioners whose focus is measuring and modelling customer loyalty and customer experience. This is a notable gap in existing research on cross-group comparisons, especially considering the reliance of business decision-makers on customer feedback. Though standard measures of customer experience and loyalty are used in every industry, their measurement invariance across industries has not been subject to extensive testing. This article brings current thinking about cross-group comparisons and modern tools of multi-group confirmatory factor analysis (MGCFA) to the measurement of customer loyalty and customer experience across firms, industries, and countries, drawing on original large-scale survey data from the United States, United Kingdom, and Canada.

Keywords: Customer research; measurement invariance; confirmatory factor analysis; United States; United Kingdom; Canada

Methods for collecting and analysing cross-national and cross-group survey data have advanced considerably in the past two decades (Harkness et al., 2010; Harkness, van de Vijver, & Mohler, 2003; Johnson, Pennell, Stoop, Ineke, & Dorer, 2019). A critical component of comparative survey research is the assessment of measurement invariance, also called measurement equivalence. Measurement invariance refers to the notion that survey-based measures capture the same underlying constructs in different groups, and thus survey estimates for these groups offer a valid basis for comparison (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014). Indeed, comparative research in the social and behavioural sciences has been seized by the question of measurement invariance. There is now a wide-ranging, multi-disciplinary literature on the cross-national measurement invariance (or non-invariance) of core values (Cieciuch, Davidov, Vecchione, Beierlein, & Schwartz, 2014; Zercher, Schmidt, Cieciuch, & Davidov, 2015), personality traits (Marsh, Nagengast, & Morin, 2013), and attitudes toward a wide range of political concepts and policy issues such as support for democracy (Ariely & Davidov, 2011), the welfare state (Stegmueller, 2011) and foreign policy attitudes (Gravelle, Reifler, & Scotto, 2017, 2020).

Existing research on measurement invariance has thus tended to focus on concepts of interest to academic sociologists, political scientists, and psychologists. Consequently, the substantive focus of this literature has little engaged researchers and practitioners focused on measuring, modelling, and comparing customer feedback, which as a seminal text on survey research methods observes, is a core application of survey research (Dillman, Smyth, & Christian, 2014, pp. 462–463). This is not to say that customer research practitioners have failed to consider the issue of measurement invariance. Existing customer research nevertheless has marked limitations, having examined only single industries (Ueltschy, Laroche, Tamilia, & Yannopoulos, 2004), single countries (Fornell, Johnson, Anderson, Cha, & Everitt Bryant, 1996; Klaus & Maklan, 2013), or considered concepts with a narrow remit such as consumer ethnocentrism (Steenkamp & Baumgartner, 1998) instead of central constructs like customer loyalty and customer experience. Indeed, the premise of long-running, cross-industry measures of customer sentiment such as the American Customer Satisfaction Index (Fornell et al., 1996) is that such inter-firm and

Acknowledgements

The author thanks Jack Chen, Sarah Cho, Jon Cohen, Austin Pettis, Caroline Queny, and Brett Silverman at SurveyMonkey for advice and assistance with survey design, translation, and data collection. They power the curious every day. The statistical analysis and discussion also greatly benefitted from the constructive comments and suggestions provided by the two anonymous reviewers.

Direct correspondence to

Timothy B. Gravelle, Senior Manager, Research Science, SurveyMonkey, Canada
E-mail: tgravelle@surveymonkey.com

inter-industry comparisons yield comparable scores and serve as valid benchmarks. Still, the validity of such measures across industries and firms is assumed rather than tested. This working assumption persists even though there is a prima facie argument for the *in*comparability of measures of customer sentiment: durable consumer goods, travel, hospitality, retail shopping and financial services imply qualitatively different (and potentially incommensurate) customer experiences. Despite this, existing research has not presented simultaneously cross-national *and* cross-industry evidence of the measurement invariance of customer sentiment using modern confirmatory factor analysis tools – the preferred approach for testing measurement invariance – to validate this working assumption (cf. Yu & Yang, 2015).

To advance the current state of customer survey research, this article brings current thinking about cross-group comparisons and modern tools of multi-group confirmatory factor analysis (MGCFA) to the measurement of customer loyalty and customer experience across countries and industries. It draws on large-scale survey data from the United States, United Kingdom, and Canada with measures of customer loyalty and experience for firms in multiple industries to assess the cross-group measurement invariance of customer sentiment, with "group" defined here as, alternately, firms, industries, and countries. In brief, it finds support for strict measurement invariance of customer loyalty and customer experience across firms (or brands) and industries, as well as across the countries studied.

Measuring Customer Loyalty and Customer Experience

The existing research literature on customer sentiment and customer behaviour is vast, offering up a veritable cacophony of competing theoretical models and empirical measures. Indeed, marketing research and management consulting firms are the same, with every firm advancing its own perspective on the optimal questions to gauge customer sentiment, and that are meant to serve as antecedents of customer behaviours of interest: customer retention, repeat purchasing, and share of wallet (e.g., Reichheld, 2003; Yu & Yang, 2015).

Customer loyalty has been described as a favourable attitude toward a brand that differentiates it from competing brands (Dick & Basu, 1994), and as the composite of beliefs, affect, and intentions toward a brand (Oliver, 1999). Defined in this way, customer loyalty refers to *attitudinal* loyalty, and is distinguishable from *behavioural* loyalty, which refers to repeat patronage or repeat purchasing (Watson, Beck, Henderson, & Palmatier, 2015). Only the former is strictly a survey-based measurement, while the latter may be measured using business operational data or as a self-reported behaviour in a survey setting (allowing for some measure-

ment error). A closely related concept, customer satisfaction, is understood as the alignment between initial customer expectations and firm performance (Fornell et al., 1996). Different approaches to customer sentiment conceive the relationship between customer satisfaction and customer loyalty differently. In some models, customer satisfaction and customer loyalty are conceived as distinct concepts (Dick & Basu, 1994; Fornell et al., 1996). Still other models – especially those current in applied customer research – subsume customer satisfaction under customer loyalty, treating satisfaction as an indicator of loyalty, along with self-reported measures of one's likelihood to recommend a brand (word-of-mouth intention) and likelihood to repurchase a brand (repurchase intention). Empirical analyses indicate that measures of customer satisfaction, likelihood to recommend, and likelihood to repurchase measure the same underlying concept (Yu & Yang, 2015).

Customer experience can be defined as customers' internal affective, cognitive, emotional, and sensorial responses to engagement with a brand (Brakus, Schmitt, & Zarantonello, 2009; Lemon & Verhoef, 2016). A variety of measures have been proposed as capturing different facets of customer experience. These include perceptions of the quality of service interactions (across touchpoints, and at different points in the customer journey), perceptions of product design and quality, value in comparison to other brands, feeling of confidence in the brand, and feeling that the brand or firm cares about one as a customer (Klaus & Maklan, 2013; Lemon & Verhoef, 2016; Schwager & Meyer, 2007; Yu & Yang, 2015). Customer experience is thus theorised as distinct from (and an antecedent of) customer loyalty (Klaus & Maklan, 2013; Lemon & Verhoef, 2016).

These theoretical considerations informed the selection of survey items intended to measure customer loyalty and customer experience. Customer loyalty is operationalised using survey items capturing customers' overall satisfaction, likelihood to repurchase, and likelihood to recommend a brand, which aligns with industry-standard measures also used by Yu and Yang (2015). Adapting existing survey items designed to measure customer experience, it is operationalised here as customers' perceptions of a brand meeting their expectations, its value for money, comparisons to other brands, the brand delivering what it promises, and how much the brand cares about them as a customer (see, e.g., Klaus & Maklan, 2013; Yu & Yang, 2015). The measurement of customer loyalty and customer experience can thus be depicted in the two-factor model shown in Figure 1.

Data

To assess the cross-firm, cross-industry, and cross-national measurement invariance of the model of customer loyalty and customer experience depicted above,

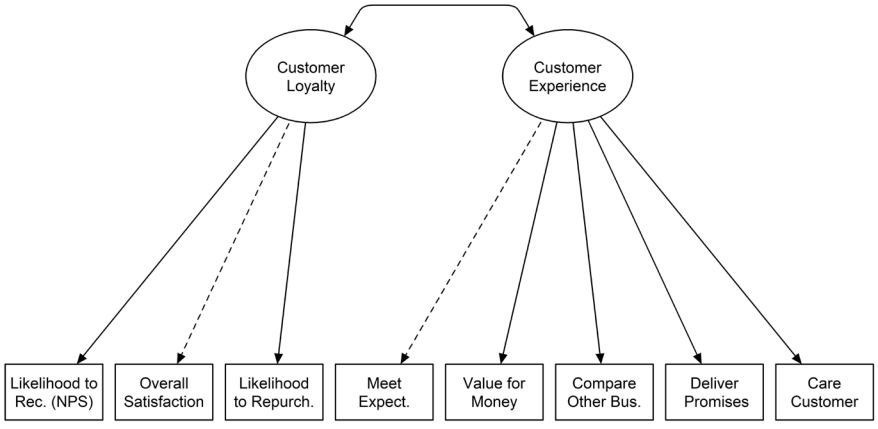


Figure 1 MGCFA Model of Customer Loyalty and Customer Experience

online surveys were conducted in the United States, United Kingdom, and Canada on 31 October–25 November 2019 using SurveyMonkey's endpage recruitment methodology. More than 2 million people around the world complete surveys designed by individuals, community organisations, and businesses using the SurveyMonkey online survey platform every day. This stream of survey respondents serves as an opportune recruitment pool for additional surveys (see, e.g., Chen, Valliant, & Elliott, 2019; Clinton, Cohen, Lapinski, & Trussler, 2021). After completing a survey on the SurveyMonkey platform, randomly selected respondents from the targeted countries (identified using their internet protocol (IP) addresses) were presented with a survey completion web page (endpage) inviting them to then complete another survey. (At the time of survey data collection, the endpage recruitment methodology was only available in these three countries.)

Five different surveys were administered in each country, each focusing on a specific industry: passenger airlines, hotels, consumer electronics, retail, and banking. These five industries represent major consumer-facing industries offering a selection of brands (or competing firms) in each national market. They also represent five of the seven industry groups represented in the ACSI: transportation, services, consumer durables, retail and finance (Fornell et al., 1996). Still, the industries and firms included in the surveys are not intended to be an exhaustive set. Respondents were first asked about their experiences (whether airline travel, hotel stays, purchases, or banking) with specific nationally leading or global companies and brands in the past 12 months. These included well-known brands such as American Airlines, British Airways, Air Canada, Hilton, Marriott, Apple, Sam-

sung, Best Buy, Wal-Mart, Bank of America, HSBC, and Royal Bank of Canada (the complete list of brands is reported in Tables A1 and A2 in the appendix). Those who reported engaging with a particular brand in the past 12 months were then asked a series of industry-standard questions measuring attitudinal customer loyalty: likelihood to recommend the brand – the widely-used "Net Promoter" question (Reichheld, 2003) – along with overall satisfaction with the brand, and likelihood to repurchase the brand. The surveys also asked about five key elements of customer experience: meeting expectations, value for money, comparisons to other brands, delivering what the brand promises, and how much the brand cares about you as a customer. All questions were asked using five-point, fully-labelled survey scales with the exception of likelihood to recommend which used the prescribed 0–10 scale, which was then recoded into the three categories specified by Reichheld (2003): "detractor" (0–6), "neutral" (7–8), and "promoter" (9–10). These eight questions are summarised in Table 1; full survey item wording appears in the appendix. The surveys were administered in English in the US and UK, and in both English and French in Canada. These samples were weighted (using weight raking) to be demographically representative of the national adult (18 years and older) populations across age, sex, region, and educational attainment categories (raking on race and ethnicity was also done in the US).

Though SurveyMonkey's endpage recruitment methodology differs from online opt-in panels, and has more in common river sampling methods, we can nevertheless calculate equivalent survey participation rates (AAPOR 2016), also called completion rates (Callegaro & Disogra, 2008), since the number of SurveyMonkey endpage views (functionally the survey invitation), click-throughs, and the number of completed surveys are all known quantities. The overall completion rate is thus calculated as 3.4 percent.

In total, 25,953 out of 41,581 respondents (or 62.4 percent) provided customer ratings, with 44,677 customer ratings collected for 60 brands across the five industries (airlines: n = 5,756; hotels: n = 6,796; consumer electronics: n = 8,347; retail: n = 16,638; banking: n = 7,140) and three countries (US: n = 12,392; UK: n = 14,974; Canada: n = 17,311). The number of ratings per brand range between 109 and 3,675, and the mean number of ratings per respondent (providing at least one brand rating) is 1.72.

It is important to acknowledge that these samples were recruited in a non-probabilistic manner. While some studies comparing probability and non-probability samples have concluded that they yield different sample point estimates (Malhotra & Krosnick, 2007; Yeager et al., 2011), other studies find few substantively meaningful differences (Ansolabehere & Schaffner, 2014; Sanders, Clarke, Stewart, & Whiteley, 2007). More germane to the present study, though, is the assessment of the measurement characteristics of a set of confirmatory factor models (more on this below) than sample point estimates for individual survey items. It is also worth

Table 1 Summary of Customer Loyalty/Experience Survey Items

Customer Loyalty
Likelihood to recommend brand (0–10, recoded into 0–6 [1], 7–8 [2], 9–10 [3])
Overall satisfaction with brand (1–5)
Likelihood to repurchase brand (industry-specific wording) (1–5)
Customer Experience
Brand met expectations (1–5)
Value for money provided by brand (1–5)
How does brand compare to other brands in industry (1–5)
How often does brand deliver what they promise (1–5)
How much does brand care about you as a customer (1–5)

noting the prevalence of non-probability samples in many marketing research and customer research applications. Given this focus on the factor structure in a customer research context, the samples collected by the endpage methodology are deemed to be fit for purpose (Baker et al., 2013).

Methods

Following from the conceptualisation of customer loyalty and customer experience presented in Figure 1 above, the models tested comprise two latent variables (or factors) corresponding to these two overarching concepts. In line with prevailing practice for testing measurement invariance, this two-factor model is analysed in a multi-group confirmatory factor analysis (MGCFA) framework (Davidov et al., 2014). As MGCFA is part of a broader structural equation modelling (SEM) framework, confirmatory factor analysis (CFA) differs from the more widely-employed exploratory factor analysis (EFA) by requiring the modeller to specify a factor model and the items measuring (i.e., that "load on") a given factor (Kline, 2016).

The typical practice for testing measurement invariance involves moving through a sequence of nested, increasingly constrained model specifications reflecting higher degrees of invariance while assessing overall model fit. *Configural invariance* is achieved when all groups have the same salient (non-zero) and non-salient (near-zero) factor loadings; no cross-group equality constraints are imposed. Configural invariance allows us to conclude that the same latent constructs exist

in all groups, but formal cross-group comparisons (e.g., of mean scores on those constructs) cannot be made. *Metric invariance* (or weak measurement invariance) requires good model fit while constraining factor loadings to be equal across groups. This allows for regression coefficients (e.g., structural relations between latent constructs) to be meaningfully compared between groups. *Scalar invariance* (or strong measurement invariance) requires good model fit while constraining factor loadings as well as item intercepts (or thresholds) to be equal across groups. This allows latent variable means meaningfully compared across groups (Davidov et al., 2014; Steenkamp & Baumgartner, 1998). Error variances (also called measurement residuals or residual variances) can be constrained to equality to test error variance invariance. *Error variance invariance* (or strict measurement invariance) allows us to conclude that a set of items serve as equally reliable indicators of the latent constructs in all groups (Steenkamp & Baumgartner, 1998). The forms of measurement equivalence described above can be subsumed under the heading of *exact* measurement invariance. Recent extensions of measurement invariance testing have investigated more flexible alternatives to exact measurement invariance, including *approximate* measurement invariance (Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014) and the *alignment method* (Asparouhov & Muthén, 2014). Given the focus here on making clear, direct comparisons of customer sentiment across countries, industries, and brands, the analysis here remains focused on exact measurement invariance.

The modelling approach pursued here thus entails testing in turn configural, metric, scalar, and scalar plus error variance invariance. Separate sets of MGCFA models are also fit where the groups comprise the 60 brands (or firms), five industries, and three countries, thus testing different levels of measurement invariance across different dimensions: brands, industries, and national contexts. Several of the ordinal survey items exhibit significant skew, with a preponderance of high scores reflecting positive brand experiences. Given the skew exhibited by the data, treating the survey data as continuous and estimating the MGCFA models by maximum likelihood is not advised (Rhemtulla, Brosseau-Liard, & Savalei, 2012). The survey data are therefore treated as categorical, and all models are estimated by robust weighted least squares (Finney & DiStefano, 2013) using M*plus* version 8.1 (Muthén & Muthén, 2017).

Concerning the assessment of overall model fit, the SEM and CFA literatures distinguish between measures of absolute fit – in particular, the model chi-square statistic – and approximate fit indices. Given the very large sample employed here, experienced structural equation modellers would expect model chi-squares to have little utility in practice: with large sample sizes, chi-square statistics routinely indicate model misfit for otherwise acceptable models. Approximate fit indices – particularly the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardised root mean square residual (SRMR) – are therefore

more useful in assessing model fit (Kline, 2016). Widely-used cut-off values for these fit indices were proposed by Hu and Bentler (1999); these are CFI $\geq$ 0.95, RMSEA $\leq$ 0.05 (or $\leq$ 0.06), and SRMR $\leq$ 0.08. An earlier proposal of Browne and Cudeck (1993) proposed RMSEA $\leq$ 0.08 as indicating acceptable model fit. Several authors, however, have argued against rigid cut-off values (regardless of the values chosen) or reliance on any single model fit index (e.g., F. Chen, Curran, Bollen, Kirby, & Paxton, 2008).

In addition to considering overall model fit, the methodological literature on measurement invariance also provides guidelines on *incremental* (or relative) model fit – that is, the change in model fit from a less constrained to a more constrained model (e.g., from configural invariance to metric invariance, or from metric invariance to scalar invariance). Initial work by Chen (2007) based on models comprising a single factor, two groups, and estimation by maximum likelihood proposed the following guidelines for the permissible change in the model fit indices: ΔCFI $\geq$ -0.01, ΔRMSEA $\leq$ 0.015, and ΔSRMR $\leq$ 0.03. Extending this work to consider multiple factors, several groups, and categorical indicators (as in the present study), Rutkowski and Svetina (2017) have proposed ΔRMSEA $\leq$ 0.05 for metric equivalence and ΔRMSEA $\leq$ 0.01 for scalar equivalence, while advising against the use of ΔCFI on account of its poor performance (their study did not examine ΔSRMR).

Accordingly, the approach to assessing model fit employed here entails examining the CFI, RMSEA, and SRMR model fit indices together to assess overall fit while also examining incremental fit, focusing on ΔRMSEA and ΔSRMR using current guidelines. In all cases, proposed cut-off values (for both overall and incremental model fit) are used as guides as opposed to rigid rules.

Results

Examining the MGCFA results for the analyses by brand, industry, and country, the model chi-square statistics (as expected) would lead one to conclude that none of the models achieve good fit (see Table 2). By contrast, the CFI, RMSEA, and SRMR model fit indices suggest good (or at a minimum, acceptable) fit.

For the MGCFA conducted with the 60 measured brands comprising the groups, the configural and metric invariant models both achieve good overall fit based on CFI ($\geq$ 0.95) and SRMR ($\leq$ 0.08); RMSEA indicates acceptable fit ($\leq$ 0.08). At a minimum, then, valid comparisons of the structural relationship between customer loyalty and customer experience can be made across firms. Still, benchmarking (i.e., comparing mean scores) on customer loyalty and customer experience metrics across firms is more typically the aim of customer research practitioners. This requires scalar invariance. The scalar invariant model again

Table 2 MGCFA Model Summary Statistics

	Model χ^2 (SB)	DF	CFI	RMSEA	(90 % c.i.)	*p* close fit	SRMR
By brand (60)							
Configural invariance	4,338.409	1,140	0.995	0.061	(0.059–0.063)	0.000	0.018
Metric invariance	4,845.274	1,494	0.995	0.055	(0.053–0.057)	0.000	0.019
Scalar invariance	10,840.100	2,674	0.988	0.064	(0.063–0.065)	0.000	0.030
Scalar + error variance invariance	12,809.688	3,138	0.986	0.064	(0.063–0.065)	0.000	0.037
By industry (5)							
Configural invariance	2,199.800	95	0.995	0.050	(0.048–0.052)	0.570	0.013
Metric invariance	2,096.521	119	0.996	0.043	(0.042–0.045)	1.000	0.013
Scalar invariance	4,139.169	199	0.991	0.047	(0.046–0.048)	1.000	0.019
Scalar + error variance invariance	4,831.688	223	0.990	0.048	(0.047–0.049)	1.000	0.023
By country (3)							
Configural invariance	1,871.708	57	0.996	0.046	(0.044–0.048)	1.000	0.011
Metric invariance	1,603.513	69	0.996	0.039	(0.037–0.040)	1.000	0.011
Scalar invariance	1,396.957	109	0.997	0.028	(0.027–0.029)	1.000	0.012
Scalar + error variance invariance	1,421.922	117	0.997	0.027	(0.026–0.029)	1.000	0.013

Note: Models are fit by robust weighted least squares. Chi-square statistics are Sartorra-Bentler scaled (mean-adjusted) chi-squares (Sartorra & Bentler, 1994).

indicates good overall fit according to the CFI (0.988) and SRMR (0.030); the RMSEA (0.064) points to increased misfit, though still yielding acceptable overall fit. The error variance invariant model is still further constrained, but still yields acceptable model fit (CFI = 0.986, RMSEA = 0.064, SRMR = 0.037). In terms of incremental model fit, the results for the metric invariance model are ΔRMSEA = -0.060 and ΔSRMR = 0.001 – well below the ΔRMSEA ≤ 0.05 and ΔSRMR ≤ 0.03 cut-offs. Change in model fit for the scalar invariance model is ΔRMSEA = 0.009 and ΔSRMR = 0.021, while change in model fit for the scalar plus error variance invariance model is ΔRMSEA = 0 and ΔSRMR = 0.007, thus meeting the ΔRMSEA ≤ 0.01 and ΔSRMR ≤ 0.03 cut-offs. Taking in the results of the overall and incremental measures of model fit, then, one can conclude that latent variable scores for customer loyalty and customer experience can be meaningfully compared across very different products and services – whether Apple or American Airlines, Best Buy or Barclays, Marriott or Marks and Spencer. Further, the three indicators of customer loyalty and five indicators of customer experience used here exhibit the same measurement properties across brands.

Not only does the MGCFA model fit across a wide variety of brands, it also fits across the five industries under study. With the groups comprised of the five industries, all model specifications – configural, metric, scalar, and scalar plus error variance invariance – achieve good overall model fit based on the guidelines for CFI (≥ 0.95), RMSEA (≤ 0.05), and SRMR (≤ 0.08) fit indices. Despite the highly constrained model specification, the error variance invariant model still indicates good model fit, with CFI = 0.990, RMSEA = 0.048, and SRMR = 0.023. The metric invariance model easily meets the guidelines for good incremental model fit with ΔRMSEA = -0.070 and ΔSRMR = 0. Change in model fit for the scalar invariance model (ΔRMSEA = 0.004, ΔSRMR = 0.006) and for the scalar plus error variance invariance model (ΔRMSEA = 0.001, ΔSRMR = 0.004) also indicate good incremental model fit. Substantively, then, latent variable scores for customer loyalty and customer experience can be compared directly across airline travel, hotel stays, consumers electronics brands, retailers, and banks.

Customer loyalty and customer experience can be similarly compared across the US, UK, and Canada. Each of the configural, metric, scalar, and scalar plus error variance invariance model exhibit good overall model fit. As with the industry groups, the highly constrained error variance invariant model achieves good model fit across countries, with CFI = 0.997, RMSEA = 0.027, and SRMR = 0.013. The metric invariance model exhibits good incremental model fit with ΔRMSEA = -0.070 and ΔSRMR = 0, as do the scalar invariant model with ΔRMSEA = -0.011 and ΔSRMR = 0.001, and the scalar plus error variance invariant model with ΔRMSEA = -0.001 and ΔSRMR = 0.001. These are important findings for large firms with global footprints and global customer bases. It implies that core customer metrics travel across the different national contexts studied. For customer

research practitioners, it similarly implies that there is little need for industry-specific customer metrics.

Looking at the standardised factor loadings for the scalar plus error variance invariance (strict measurement invariance) MGCFA models, it is worth pointing out that they are consistently high (see Table 3). Across groups by brand, industry, and country, and across all items, factor loadings range between 0.757 and 0.954. The items thus serve as good measures of customer loyalty and customer experience, respectively. It is also worth noting the standardised correlations between customer loyalty and customer experience factors, which range between 0.831 and 0.992 for specific brands; slightly narrower ranges are seen for the industry- and country-grouped models. These high correlations might suggest a lack of discriminant validity to readers accustomed to lower factor correlations, though such high correlations are common in customer research (e.g., Fornell et al., 1996). More to the point, these data pass the conventional CFA test of discriminant validity where the factor correlation is constrained to be equal to 1, implying a one-factor model. This test is highly significant ($\chi^2 = 2{,}069.515$, d.f. $= 1$, $p < 0.001$), indicating that a one-factor model has significantly worse fit than a two-factor model. A two-factor model thus remains preferable despite the high standardised correlation between the customer loyalty and customer experience factors.

Conclusion

The question motivating this article was whether widely used measures of customer loyalty and customer experience translate across, firms (or brands), industries, and countries. This question has immense practical importance for firms seeking to win and retain customers, to benchmark their performance against competitors, or to benchmark their performance in different markets. Without a rigorous basis for comparisons of customer loyalty and customer experience across competing brands, or comparisons of brand performance across countries – that is, without measurement invariance – one could truly be relying on an apples-to-oranges comparison to make critical business decisions. Indeed, a great deal of applied customer research is premised on the comparability of survey-based measures of customer sentiment, even though airline flights, hotel stays, smartphones, retail shopping, and everyday banking imply qualitatively different experiences.

The goal of this article, then, was to advance the literatures on customer research and consumer behaviour (as well as the literature on cross-national survey research more broadly) by presenting the first large-scale study of cross-firm, cross-industry, and cross-national measurement invariance of customer sentiment using MGCFA tools. The analyses presented here indicate that rigorous quantitative comparisons are in fact well-grounded. Whether examined across brands,

Table 3 MGCFA Standardised Factor Solutions (Scalar + Error Variance Invariance)

	By Brand (60)		By Industry (5)		By Country (3)	
	Customer Loyalty	Customer Experience	Customer Loyalty	Customer Experience	Customer Loyalty	Customer Experience
Factor Loadings						
Likelihood to recommend	0.829–0.921	–	0.846–0.898	–	0.861–0.881	–
Overall satisfaction	0.880–0.954	–	0.899–0.938	–	0.903–0.923	–
Likelihood to repurchase	0.828–0.920	–	0.843–0.892	–	0.855–0.876	–
Meet expectations	–	0.860–0.938	–	0.878–0.911	–	0.880–0.896
Value for money	–	0.770–0.921	–	0.830–0.881	–	0.827–0.853
Compare to other brands	–	0.846–0.931	–	0.866–0.901	–	0.862–0.884
Deliver what they promise	–	0.798–0.905	–	0.829–0.871	–	0.829–0.858
Care about you as a customer	–	0.757–0.903	–	0.814–0.866	–	0.805–0.836
Factor Means	-1.697–0.000	-1.396–0.110	0.000–0.554	0.000–0.549	-0.039–0.191	-0.043–0.197
Factor Correlation						
Customer Loyalty – Customer Experience	0.831–0.992		0.856–0.948		0.915–0.923	

Note: Factor loadings are standardised. Factor loadings for overall satisfaction (factor 1 – Customer Loyalty) and care about you as a customer (factor 2 – Customer Experience) are fixed to 1 in the unstandardised solution for model identification. Likelihood to recommend is recoded from 0–10 to 1–3.

industries, or countries, a simple two-factor model comprising customer loyalty and customer experience exhibits scalar plus error variance invariance (strict measurement invariance), providing a firm basis for cross-group comparisons. These results should be welcomed by customer research practitioners, since they imply that industry-standard measures of customer sentiment exhibit robust measurement properties.

These results should nevertheless be interpreted in light of the finite number of industries examined. Extending this research to other industries – for example, the automotive sector, insurance, computer software, restaurants, and consumer packaged goods, among others – would assist in reconfirming or qualifying the findings presented here. Similarly, the number of countries included in the analyses is finite. In particular, it is important to acknowledge that the US, UK, and Canada all comprise English-speaking majorities, meaning this study has largely set aside the question of cross-language measurement invariance (but see Yu & Yang, 2015). Testing the measurement invariance of customer loyalty and customer experience across a larger number of countries and languages, as others have done on other substantive topics (e.g., Davidov & De Beuckelaer, 2010; Gravelle et al., 2017, 2020), would be a valuable test of the model advanced here.

More broadly, this article argues that survey researchers engaged in applied customer research – perhaps employed as an in-house analyst charged with assessing their firm's position in the marketplace vis-à-vis competing brands, or as a marketing researcher consulting to a large firm with a global customer base – *should* be concerned with the question of cross-group measurement invariance, and ought to examine it explicitly instead of leaving it as an untested assumption. This article provides a demonstration of how to do so using current MGCFA techniques.

References

American Association for Public Opinion Research. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.). AAPOR.

Ansolabehere, S., & Schaffner, B. F. (2014). Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison. *Political Analysis*, *22*(3), 285–303.

Ariely, G., & Davidov, E. (2011). Can we Rate Public Support for Democracy in a Comparable Way? Cross-National Equivalence of Democratic Attitudes in the World Value Survey. *Social Indicators Research*, *104*(2), 271–286.

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling*, *21*(4), 495–508.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., … Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–105.

Brakus, J. J., Schmitt, B. H., & Zarantonello, L. (2009). Brand Experience: What Is It? How Is It Measured? Does It Affect Loyalty? *Journal of Marketing*, *73*(3), 52–68.

Browne, M. W., & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136–162). Newbury Park, CA: Sage.

Callegaro, M., & Disogra, C. (2008). Computing Response Metrics for Online Panels. *Public Opinion Quarterly*, *72*(5), 1008–1032.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling*, *14*(3), 464–504.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models. *Sociological Methods and Research*, *36*(4), 462–494.

Chen, J. K. T., Valliant, R. L., & Elliott, M. R. (2019). Calibrating Non-Probability Surveys to Estimated Control Totals Using LASSO, with an Application to Political Polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3), 657–681.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology*, *5*(982), 1–10.

Cieciuch, J., Davidov, E., Vecchione, M., Beierlein, C., & Schwartz, S. H. (2014). The Cross-National Invariance Properties of a New Scale to Measure 19 Basic Human Values: A Test Across Eight Countries. *Journal of Cross-Cultural Psychology*, *45*(5), 764–776.

Clinton, J., Cohen, J., Lapinski, J. S., & Trussler, M. (2021). Partisan Pandemic: How Partisanship and Public Health Concerns Affect Individuals' Social Distancing During COVID-19. *Science Advances*, 7(2), eabd7204.

Davidov, E., & De Beuckelaer, A. (2010). How Harmful are Survey Translations? A Test with Schwartz's Human Values Instrument. *International Journal of Public Opinion Research*, *22*(4), 485–510.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, *40*(1), 55–75.

Dick, A. S., & Basu, K. (1994). Customer Loyalty: Toward an Integrated Conceptual Framework. *Journal of the Academy of Marketing Science*, *22*(2), 99–113.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th ed.). Hoboken, NJ: Wiley,

Finney, S. J., & DiStefano, C. (2013). Non-Normal and Categorical Data in Structural Equation Modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (2nd ed., pp. 439–492). Charlotte, NC: Information Age Publishing.

Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Everitt Bryant, B. (1996). The American Customer Satisfaction Index: Nature, Purpose, and Findings. *Journal of Marketing*, *60*(4), 7–18.

Gravelle, T. B., Reifler, J., & Scotto, T. J. (2017). The Structure of Foreign Policy Attitudes in Transatlantic Perspective: Comparing the United States, United Kingdom, France and Germany. *European Journal of Political Research*, *56*(4), 757–776.

Gravelle, T. B., Reifler, J., & Scotto, T. J. (2020). The Structure of Foreign Policy Attitudes among Middle Power Publics: A Transpacific Replication. *Australian Journal of International Affairs*.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., … Smith, T. W. (Eds.). (2010). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. New York, NY: Wiley.

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. P. (Eds.). (2003). *Cross-Cultural Survey Methods*. New York, NY: Wiley.

Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Johnson, T. P., Pennell, B.-E., Stoop, Ineke, A. L., & Dorer, B. (Eds.). (2019). *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*. Hoboken, NJ: Wiley.

Klaus, P., & Maklan, S. (2013). Towards a better measure of customer experience. *International Journal of Market Research*, *55*(2), 227–246.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4th ed.). New York, NY: Guilford.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey. *Journal of Marketing*, *80*(6), 69–96.

Malhotra, N., & Krosnick, J. A. (2007). The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples. *Political Analysis*, *15*(3), 286–323.

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*, *49*(6), 1194–1218.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Oliver, R. L. (1999). Whence Consumer Loyalty? *Journal of Marketing*, *63*(S1), 33–44.

Reichheld, F. F. (2003). The One Number You Need to Grow. *Harvard Business Review*, (December), 1–10.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.

Rutkowski, L., & Svetina, D. (2017). Measurement Invariance in International Surveys: Categorical Indicators and Fit Measure Performance. *Applied Measurement in Education*, *30*(1), 39–51.

Sanders, D., Clarke, H. D., Stewart, M. C., & Whiteley, P. (2007). Does Mode Matter For Modeling Political Choice? Evidence From the 2005 British Election Study. *Political Analysis*, *15*(3), 257–285.

Sartorra, A., & Bentler, P. M. (1994). Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent Variables Analysis: Applications for Developmental Research* (pp. 399–419). Thousand Oaks, CA: Sage.

Schwager, A., & Meyer, C. (2007). Understanding Customer Experience. *Harvard Business Review*, (February), 116–128.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research*, *25*(1), 78–90.

Stegmueller, D. (2011). Apples and Oranges? The Problem of Equivalence in Comparative Research. *Political Analysis*, *19*(4), 471–487.

Ueltschy, L. C., Laroche, M., Tamilia, R. D., & Yannopoulos, P. (2004). Cross-cultural invariance of measures of satisfaction and service quality. *Journal of Business Research*, *57*(8), 901–912.

Watson, G. F., Beck, J. T., Henderson, C. M., & Palmatier, R. W. (2015). Building, measuring, and profiting from customer loyalty. *Journal of the Academy of Marketing Science*, *43*(6), 790–825.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly*, *75*(4), 709–747.

Yu, D., & Yang, Y. (2015). Measurement Equivalence of a Concise Customer Engagement Metric across Country, Language, and Customer Types. *Public Opinion Quarterly*, *79*(S1), 325–358.

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance. *Frontiers in Psychology*, *6*(733), 1–11.

Appendix A:

Survey Questionnaires

[AIRLINES:] In the past 12 months, which of the following airlines have you traveled on? (Please select all that apply.)

[HOTELS:] In the past 12 months, which of the following hotel chains have you stayed at? (Please select all that apply.)

[CONSUMER ELECTRONICS:] In the past 12 months, have you purchased any consumer electronics (for example, a television, desktop computer, laptop, tablet, smartphone, wearable device) from any of the following brands? (Please select all that apply.)

[RETAIL:] In the past 12 months, which of the following stores have you shopped at? (Please select all that apply.)

[BANKING:] In the past 12 months, have you done any banking (through a checking account, savings account, mortgage, or personal line of credit) with any of the following banks or financial institutions? (Please select all that apply.)

How likely is it that you would recommend [BRAND] to a friend or colleague?

0 – Not at all likely
1
2
3
4
5
6
7
8
9
10 – Extremely likely

Overall, how satisfied or dissatisfied are you with [BRAND]?

Very satisfied
Somewhat satisfied
Neither satisfied nor dissatisfied
Somewhat dissatisfied
Very dissatisfied

How likely are you to [AIRLINES: travel with BRAND again] [HOTELS: stay at BRAND again] [ELECTRONICS: purchase BRAND products again] [RETAIL: shop at BRAND again] [BANKING: continue to bank with BRAND]?

- Extremely likely
- Very likely
- Moderately likely
- Slightly likely
- Not at all likely

How well have your experiences with [BRAND] met your expectations?

- Much better than expected
- Better than expected
- About what I expected
- Worse than expected
- Much worse than expected

How would you rate the value for money provided by [BRAND]?

- Excellent
- Above average
- Average
- Below average
- Poor

How does [BRAND] compare to other [AIRLINES: airlines] [HOTELS: hotel companies] [ELECTRONICS: consumer electronics companies] [RETAIL: stores] [BANKING: banks]? Are they...?

- Much better
- Better
- About the same
- Worse
- Much worse

How often does [BRAND] deliver what they promise?

- Always
- Usually
- Sometimes
- Rarely
- Never

How much does [BRAND] care about you as a customer?

- A great deal
- A lot
- A moderate amount
- A little
- Not at all

Appendix B:

Brands Rated by Industry and Country

	United States (n = 12,392)	United Kingdom (n = 14,974)	Canada (n = 17,311)
Airlines (n = 5,756)	United States (n = 1,272)	United Kingdom (n = 1,915)	Canada (n = 2.569)
	American Airlines Delta Airlines JetBlue Airways Southwest Airlines United Airlines	British Airways EasyJet Jet2 RyanAir TUI Airways	Air Canada Air Transat Porter Airlines WestJet
Hotels (n = 6,796)	United States (n = 1,489)	United Kingdom (n = 2,126)	Canada (n = 3,181)
	Best Western Comfort Inn Hilton Holiday Inn Marriott Travelodge	Best Western Hilton Holiday Inn Marriott Premier Inn Travelodge	Best Western Comfort Inn Hilton Holiday Inn Marriott Travelodge
Consumer electronics (n = 8,347)	United States (n = 2,168)	United Kingdom (n = 3,785)	Canada (n = 2,394)
	Apple Fitbit LG Microsoft Samsung Sony	Apple Fitbit LG Microsoft Samsung Sony	Apple Fitbit LG Microsoft Samsung Sony

	United States (n = 12,392)	United Kingdom (n = 14,974)	Canada (n = 17,311)
Retail (n = 16,638)	United States (n = 5,017)	United Kingdom (n = 5,192)	Canada (n = 6,429)
	Best Buy Costco Macy's Target The Home Depot Wal-Mart	Asda Currys PC World Homebase Marks & Spencer Tesco	Best Buy Canadian Tire Costco Hudson's Bay The Home Depot Wal-Mart
Banking (n = 7,140)	United States (n = 2,446)	United Kingdom (n = 1,956)	Canada (n = 2,738)
	Bank of America BB&T Capital One Bank Chase Bank PNC Bank U.S. Bank Wells Fargo	Barclays Halifax HSBC Lloyds TSB Nationwide NatWest Royal Bank of Scotland Santander	BMO Bank of Montreal CIBC HSBC Royal Bank of Canada Scotiabank Tangerine TD Canada Trust

Appendix C:

Ratings by Brand

Brand	Countries	n =	Brand	Countries	n =
Apple	US, UK, Canada	3,675	American Airlines	US	466
Wal-Mart	US, Canada	3,195	Delta Airlines	US	466
Samsung	US, UK, Canada	2,287	Wells Fargo	US	448
The Home Depot	US, Canada	2,083	Scotiabank	Canada	440
Costco	US, Canada	1,783	CIBC	Canada	416
Tesco	UK	1,776	Barclays	UK	415
Marriott	US, UK, Canada	1,531	Macy's	US	403
Asda	UK	1,401	EasyJet	UK	393
Air Canada	Canada	1,339	Sony	US, UK, Canada	370
Hilton	US, UK, Canada	1,296	Santander	UK	370
Best Buy	US, Canada	1,291	United Airlines	US	362
Holiday Inn	US, UK, Canada	1,167	Lloyds TSB	UK	347
Canadian Tire	Canada	1,136	BMO Bank of Montreal	Canada	346
Target	US	1,078	NatWest	UK	327
Best Western	US, UK, Canada	1,047	Nationwide	UK	325
Marks & Spencer	UK	1,030	Homebase	UK	316
WestJet	Canada	884	HSBC	UK, Canada	316

Brand	Countries	n =	Brand	Countries	n =
TD Canada Trust	Canada	733	British Airways	UK	300
Microsoft	US, UK, Canada	711	Capital One Bank	US	286
Comfort Inn	US, Canada	681	RyanAir	UK	274
LG	US, UK, Canada	663	Halifax	UK	268
Hudson's Bay	Canada	652	Air Transat	Canada	222
Fitbit	US, UK, Canada	641	Jet2	UK	170
Royal Bank of Canada	Canada	630	Tangerine	Canada	142
Premier Inn	UK	548	U.S. Bank	US	136
Travelodge	US, UK, Canada	526	TUI Airways	UK	135
Southwest Airlines	US	506	PNC Bank	US	129
Currys PC World	UK	494	Porter Airlines	Canada	124
Bank of America	US	480	JetBlue Airways	US	115
Chase Bank	US	477	Royal Bank of Scotland	UK	109

methods, data, analyses | 2021, Vol. 15(2), pp. 215-240 DOI: 10.12758/mda.2021.04

A Critical Evaluation of Tracking Public Opinion with Social Media: A Case Study in Presidential Approval

Robyn A. Ferg[1], *Frederick G. Conrad*[2] & *Johann A. Gagnon-Bartsch*[1]

[1]*University of Michigan Department of Statistics*

[2]*University of Michigan Program in Survey Methodology*

Abstract

There has been much interest in using social media to track public opinion. We introduce a higher level of scrutiny to these types of analyses, specifically looking at the relationship between presidential approval and "Trump" tweets and developing a framework to interpret its strength. We use placebo analyses, performing the same analysis but with tweets assumed to be unrelated to presidential approval, to assess the relationship and conclude that the relationship is less strong than it might otherwise seem. Secondly, we suggest following users longitudinally, which enables us to find evidence of a political signal around the 2016 presidential election. For the goal of supplementing traditional surveys with social media data, our results are encouraging, but cautionary.

Keywords: social media, Twitter, surveys, sentiment analysis, presidential approval

Surveys are critical for understanding public opinion and setting public policy. While asking survey questions to samples designed to represent the entire population has been very successful for many years, surveys are becoming increasingly costly to perform and response rates are declining (e.g. de Leeuw and de Heer (2002)). One proposed alternative to traditional surveys, as laid out by the AAPOR task force on big data (Murphy, et al., 2014), is to use data gathered from social media to supplement or in some cases replace traditional surveys (Hsieh & Murphy, 2017).

Early analyses were promising, finding high correlations when tracking public opinion surveys with tweets containing a given keyword. For example, O'Connor, Balasubramanyan, Routledge, & Smith (2010) found high correlations between sentiment of tweets from 2008-2009 containing the word "jobs" and survey-based measures of consumer confidence, as well as a high correlation between the sentiment of tweets from 2009 containing the word "Obama" and survey-based measures of presidential approval. Cody, Reagan, Dodds, & Danforth (2016) found similar correlations using more recent tweets through 2015. Daas & Puts (2014) found high correlations between sentiment of various subsets of Dutch social media messages and consumer confidence in the Netherlands. These findings suggest there may be an underlying relationship between data extracted from social media and public opinion surveys.

However, inconsistencies in these initial analyses warrant skepticism in underlying relationships between social media data and survey responses. In O'Connor et al. (2010), a high correlation is observed between Obama's standing in 2008 presidential election polls and the frequency---but not sentiment---of "Obama" tweets. Surprisingly, however, O'Connor et al. (2010) also found a positive correlation between Obama's standing in election polls and the frequency of tweets that contain the word "McCain". O'Connor et al. (2010) did not find a relationship between "job" (as opposed to "jobs") tweets or "economy" tweets and consumer confidence, raising concerns about the robustness of the findings. Further confusing this issue, Cody et al. (2016) did find a relationship between "job" tweets and consumer confidence, resulting in a set of subtly contradictory findings. Daas & Puts (2014) found correlations between Dutch consumer sentiment and various subsets of Dutch social media messages (such as messages containing pronouns, messages containing the most frequent spoken and written words in Dutch, and messages containing

Acknowledgements

This work was supported by the National Science Foundation Division of Mathematical Sciences [1646108 to R.A.F.].

Direct correspondence to

Robyn A. Ferg, University of Michigan Department of Statistics, West Hall, 1085 S. University Ave., Ann Arbor, MI 48109
E-mail: fergr@umich.edu

the Dutch equivalents of "the" and "a/an") that were just as strong as messages containing words about the economy, raising red flags for whether the economic tweets were truly capturing consumer confidence.

Upon further analysis, the initial relationships that appear strong between Twitter data and public opinion surveys can easily fall apart. Conrad et al. (2019) further investigated the relationship between sentiment of "jobs" tweets and consumer confidence, finding that seemingly small changes in sentiment calculation can drastically change the strength of the resulting relationship. Neither sorting "jobs" tweets into various categories (e.g. news/politics, job advertisements) (Conrad, et al., 2019) nor weighting survey responses to reflect the population of Twitter users (Pasek, Yan, Conrad, Newport, & Marken, 2018) restored the relationship. Furthermore, correlations between sentiment of "jobs" tweets and consumer confidence were found to be unstable over time (Conrad, et al., 2019; Pasek, et al., 2018). Conrad et. al. concluded that correlations between consumer confidence and sentiment of "jobs" tweets as reported in O'Connor et al. were likely spurious.

With the benefit of hindsight, it is perhaps not surprising that public opinion for select topics, such as the economy, can be difficult to obtain from social media. For example, even if a user's "jobs" tweet is about the economy (as opposed to, for example, Steve Jobs), the user's opinion about the economy is not always clear from the tweet. Tweets about politics, on the other hand, are often quite clear with regard to who or what a user supports or opposes. Therefore, if there is a strong, reliable signal present in Twitter that might be used to supplement traditional surveys, we might reasonably expect to find it in the political realm. In addition, there is some evidence that non-probability online survey panels produce plausible estimates of Americans' political affiliation and ideology, despite very different sampling practices. Kennedy et al. (2016) compared the estimates of political affiliation and ideology derived from responses to a questionnaire administered to samples from nine non-probability panels. All told essentially the same story about political affiliation (all somewhat overestimated the proportion of Democrats and somewhat underestimated the proportion of Independents) and ideology (Democrats were likely to favor a government that does more, within seven points of a gold standard based on telephone surveys of representative samples, and Republicans were likely to believe the government does too many things, within eight points of the gold standard). For these reasons, we focus our attention in this paper on tracking presidential approval, which we regard as "best-case scenario" for the goal of using social media data to supplement traditional surveys.

There are two main contributions in this paper. Our first contribution is methodological. If social media are to be reliably used to track public opinion, there needs to be a method of evaluating the strength of associations between social media data and public opinion surveys. While the results of Conrad et al. (2019) and Pasek et al. (2018) cast doubt on the credibility of previously observed rela-

tionships between Twitter sentiment and public opinion surveys, there remains a need for a systematic framework to interpret the strength of such relationships. To address this we propose the use of *placebo analyses*. The idea behind a placebo analysis is to replicate the primary analysis but using variables that are known to have no true relationship with the response. As an example of a placebo analysis, DiNardo & Pischke (1996) revisited a previous study that claimed wage differentials were due to computer use in the workplace. When replacing the variable for computer use in the analysis with pen/pencil use, the estimated effect of pencil use on wage differentials was similar to the estimated effect of computer use. This casts doubt on the original claim that computers in the workplace were causing the wage differential since the true effect for the placebo variable (pencil use) should be zero. The implication of an estimated non-zero effect is that the original analysis was not credible, see Athey & Imbens (2017) for further details. We develop a framework to evaluate and interpret the strength of observed correlations between social media sentiment and public opinion surveys by essentially performing multiple placebo tests. In the context of presidential approval, we first calculate the correlation between survey-based measures of presidential approval and the sentiment of tweets that contain the word "Trump". In doing so, however, we adjust smoothing and lag parameters to obtain the best possible correlation, as is typically done in similar analyses (Conrad et al. 2019, O'Connor et al. 2010). Because we optimize over these parameters, it is difficult to interpret the strength of the resulting correlation. We therefore compare our observed correlation to other correlations that are calculated in a similar way, but which are assumed to be spurious. Using this framework, we conclude that while there may be a signal when tracking sentiment of tweets containing the word "Trump", it is small and not obviously useful. These results cast doubt on whether Twitter data can reliably be used as a replacement for traditional surveys.

Our second contribution deals with the method in which social media data are obtained. As an alternative to the commonly used method of simply collecting tweets that contain a given keyword (e.g., "Trump") irrespective of who is posting them, we propose following a set of politically active Twitter users over time. This method of collecting tweets is similar to Golder & Macy (2011), who tracked mood using up to 400 tweets for each of millions of users. By collecting tweets in this manner we can track changes in sentiment among a fixed set of users. We classify politically active Twitter users as a Democrat or Republican and find evidence of a political signal when tracking both the frequency and sentiment of these users' tweets around the 2016 U.S. presidential election.

Relationship Between "Trump" Tweets and Presidential Approval

We obtain survey based measures of presidential approval from the website FiveThirtyEight.com, which aggregates multiple presidential approval surveys and weights each survey by sample size and pollster quality rating (based on historical accuracy in predicting election results and methodological standards) to obtain an overall measure of daily presidential approval (Silver, 2017).

We scrape 1000 tweets per day containing the word "Trump" during the time period from January 20, 2017 through August 25, 2019. This particular interval started with the first day of the Trump administration and covered the following 31 months. Sentiment of individual tweets is calculated using Vader, a rule-based sentiment method trained on tweets and shown to perform well at assessing sentiment of tweets (Hutto & Gilbert, 2014). Vader assigns a continuous sentiment score between -1 and 1 to each individual tweet. Vader takes into account multiple lexical features of the tweets (e.g. capitalization, punctuation, emojis), and therefore it was not necessary to perform any text cleaning of the tweets.

We do not have access to individual presidential approval survey responses nor do we know the actual political opinions of each of the users that appear in our sample of 1000 "Trump" tweets per day. Therefore, we cannot perform linkage at an individual level, as is often done in political communication studies (De Vreese, Boukes, Schuck, Vliegenthart, Bos & Lelkes, 2017). Instead, we search for an aggregate-level relationship between daily presidential approval and daily sentiment of "Trump" tweets over the given time period.

There is much variation in mean Twitter sentiment day-to-day. This variation is intrinsic to Twitter (that is, it cannot be simply attributed to our limited sampling of 1000 tweets per day; see Appendix A for details). To address this daily variation, we introduce a smoothing parameter k: the smoothed Twitter sentiment for a given day is calculated by taking the average sentiment of that day and previous k-1 days. We also introduce a lag term L, shifting survey responses ahead or behind by L days. This tells us whether Twitter sentiment leads or lags presidential approval. We allow k to be in {1, 2,…,45} and L to be in {-30, -29,…,29, 30}. We choose k and L such that we obtain the highest correlation between sentiment of "Trump" tweets and presidential approval. We choose k and L in this manner for three reasons: (1) it is not clear a priori whether social media lags survey responses or vice versa and it is not clear what the optimal smoothing might be, (2) we want to give the political signal the best chance of emerging, and (3) similar methods were performed in previous analyses (e.g. O'Connor et al. (2010) and Cody et al. (2016)). An optimal smoothing of 45 days and lag of 30 days (meaning that Twitter sentiment lags presidential approval by 30 days) gives the maximum correlation of 0.516 between sentiment of "Trump" tweets and presidential approval. While this is

not as high as previously observed correlations between "Obama" tweets and presidential approval (0.73 in O'Connor et al. (2010) and 0.76 in Cody et al. (2016)), the correlation of 0.516 might still seem to suggest there is a relationship between sentiment of "Trump" tweets and presidential approval from 2017 through mid-2019.

The observed correlation of 0.516 appears to be moderately strong. However, we optimized over the smoothing and lag parameters, and trends in time-series data can artificially inflate correlations, so it is unclear how to interpret the strength of the 0.516 correlation. To accurately interpret the strength of this observed correlation, we want to know how large the correlation would be if there were no underlying relationship between "Trump" tweets and presidential approval. To do this, we use a random sample of 5000 tweets per day from the same time frame. We first extract all words and symbols (such as emojis and numbers) that appear in at least one tweet per day in this data set. After removing stop words (e.g. "the", "an"), we are left with 497 words and symbols. We call these placebo words, as the only relationships between sentiment of tweets containing a given placebo word and presidential approval are presumably spurious. There are some "Trump" tweets in our random sample of all tweets, but they constitute a small percentage of our random sample. For each of these placebo words we repeat the same analysis as we did with the "Trump" tweets. That is, using tweets that contain a given placebo word, we adjust smoothing and lag such that we obtain the maximum absolute correlation between sentiment of tweets containing the placebo word and presidential approval. Due to the method in which placebo words are extracted, the daily sample size of tweets varies from day to day and is often less than the 1000 tweets per day as with the "Trump" tweets. Further discussion of optimal smoothing and lag parameters is given in Online Appendix B. This results in 497 placebo correlations. We call the set of these correlations the reference distribution. Figure 1 gives the reference distribution. The reference distribution is bimodal. This is because we manipulate the smoothing and lag parameters to find the optimal correlation (in absolute value) between sentiment of tweets containing each of the placebo words and presidential approval. To assess the strength of the relationship between "Trump" tweets and presidential approval, we compare the observed correlation in relation to the reference distribution. If there truly is a relationship between sentiment of "Trump" tweets and presidential approval, the observed correlation should be much larger than nearly all of the placebo correlations. Our observed correlation of 0.516 is represented by the dashed vertical line in Figure 1 and is larger than many of the placebo correlations, but not considerably so. About 4.6% of the placebo correlations are larger in absolute value than the correlation between presidential approval and "Trump" tweets (see Online Appendix B for further details). However, none of the placebo words with maximum absolute correlations greater than 0.516 are meaningfully related to presidential approval, e.g., "wanted", "tweet", "enough", "17", and "000" are five of the top words with the highest maximum absolute cor-

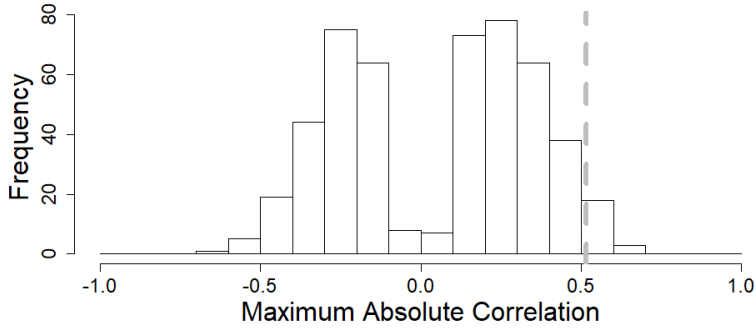


Figure 1 Reference distribution of maximum absolute correlations between presidential approval and sentiment of 497 placebo words with k in {1,…,45} and L in {-30, -29,…, 29, 30}, with bin widths of 0.1. Maximum correlation between sentiment of "Trump" tweets and presidential approval, 0.516, is denoted by the vertical dashed line.

relation with presidential approval. While there appears to potentially be a signal, if anything it is a very weak signal, and a signal that is not significantly stronger than ones found with a random sample of tweets unrelated to politics.

Note that this placebo analysis framework can be used to evaluate the strength of any measure of association and any pre-processing of sentiment between messages containing some keyword and survey responses, not just correlation when adjusting for smoothing and lag in the context of presidential approval.

Longitudinal Analysis of Twitter Users

The results of the previous section raise concern on the utility of tracking public opinion with tweets that contain a given word over time. This is not an encouraging result, suggesting that it may not typically be possible to recover strong, non-spurious alignment between survey responses and Twitter data in this manner. Indeed, alignment between survey responses and social media data is rare and nontrivial, as demonstrated by the findings reported in the previous section and by seemingly strong relationships not holding up over time (e.g. Conrad et al. (2019) replicated key findings in O'Connor et al. (2010) in the original time period but were unable to detect alignment after that). However, we believe the jury is still out on the usefulness of social media data in tracking public opinion over longer time scales. It has been observed that Twitter reacts to the onset of events on short term time scales

(Pasek, McClain, Newport, & Marken, 2019), but we are interested in longer term trends in public opinion. Our goal here is to further investigate whether Twitter may indeed contain valuable information for the purpose of tracking long term trends in public opinion, and if so, how it might be better identified.

In this section we propose an alternative approach: instead of tracking tweets containing a given word (e.g. "Trump"), we follow a group of users longitudinally. A longitudinal study of Twitter users performed in this manner may have several advantages. For example, when following the word "Trump" over time, we cannot be sure as to what extent the demographics of users tweeting about Trump are changing over time. By holding the set of users constant, we remedy this issue. Our goal in this section is to detect some aspect of the data that is clearly related to the political feelings of the set of Twitter users and is convincingly non-spurious. Note that unlike in the previous section, our goal is not to find a relationship between data extracted from Twitter and general public opinion survey responses. Instead, we examine tweets for a set of Twitter users around what we assume to be one of the most consequential events to occur on Twitter for this set of users: the outcome of the 2016 presidential election.

Similar to the previous section, we attempt to choose a setting in which the signal has the best chance of emerging. We first gather an appropriate set of Twitter users, i.e., a set of politically active users. We define a user as politically active if their location was determined to be within the United States and they produced at least 20 original (non-retweet) tweets in 2016, at least 10 of which were political (determined by whether a tweet contained at least one word from a hand-created list of political words). We had a total of 4189 politically active users. See Online Appendix C for further details on gathering our set of politically active users.

Since we are tracking a political signal and members of different parties often have opposing views regarding the lead up to and outcome of the 2016 election, we would ideally like to know each user's political party affiliation. While it can be difficult to determine political affiliation of users who are not politically engaged on Twitter (Cohen and Ruths, 2013), we are specifically considering users that are at least minimally politically active. We create a training set of users with known political affiliation, Democrat or Republican, by hand-classifying users whose self-provided profile description contained a political word. Our training set consisted of 170 Democrats and 393 Republicans. Using this set of users we build a classifier to predict political affiliation of the remaining users. Previous studies that have classified Twitter users into political party often rely on users' posts and other profile information such as name, self-reported location, and profile picture (e.g. Conover, Gonçalves, Ratkiewicz, Flammini & Menczer, 2011; Vijayaraghavan,, Vosoughi & Roy, 2017; Pennacchiotti & Popescu, 2011). In our approach we focus on the following network of each of our politically active users. As covariates for the classifier we used the list of accounts that at least 30 of the users with known political

affiliation follow. There are 3040 such accounts. A random forest is used as the classifier. The random forest appears to perform well, with only 2.66% of users with known political party being incorrectly classified and the most important accounts for classification being either politicians, political commentators, or family members of politicians. A confusion matrix and variable importance plot can be found in Appendix C. We use the trained random forest to predict political party for the remaining politically active users with unknown political party and apply an 80% cutoff rate (meaning a user is classified as a member of a given political party if at least 80% of the trees predict the user to be a member of that party), which gives 489 total Democrats and 996 total Republicans that we use going forward. There are over twice as many Republicans as Democrats in this set of users. This could potentially be for two reasons: (1) our politically active users came from a data set of tweet containing the word "jobs", and Republicans may be more likely to tweet about "jobs" compared to Democrats, or (2) Democrats are slightly more difficult to classify, so the uneven split may be due to the 80% cutoff rate. See Appendix C for further details.

We consider two metrics for tracking the tweets of our set of Democratic and Republican users: frequency and sentiment. Frequency tells whether or not our set of users are tweeting about political events, and sentiment tells us their reaction to those events. These two metrics are adjusted for the number of users in each party, so despite the uneven split between Democrats and Republicans the metrics are directly comparable between parties. We first consider the frequency of all original (i.e., non-retweet) tweets sent by our set of Democratic and Republican Twitter users. Figure 2 shows the frequency of original tweets for Democrats and Republicans from 2016 through mid-2017. The solid vertical lines on these plots represent election day (November 8, 2016) and inauguration day (January 20, 2017) and the dashed vertical lines represent the top four days with the highest frequency of tweets. The top four days with the highest frequency of tweets for Democrats, in order of frequency, are October 10, 2016; November 9, 2016; October 20, 2016; and September 27, 2016. These days correspond to the day after the election and the days after the three presidential debates between Hillary Clinton and Donald Trump. The top four days for Republicans are November 9, 2016; October 20, 2016; October 10, 2016; and November 8, 2016. These days correspond to the day after the election, days after the third and second debates, and election day. The frequency of tweets is clearly politically driven for both Democrats and Republicans.

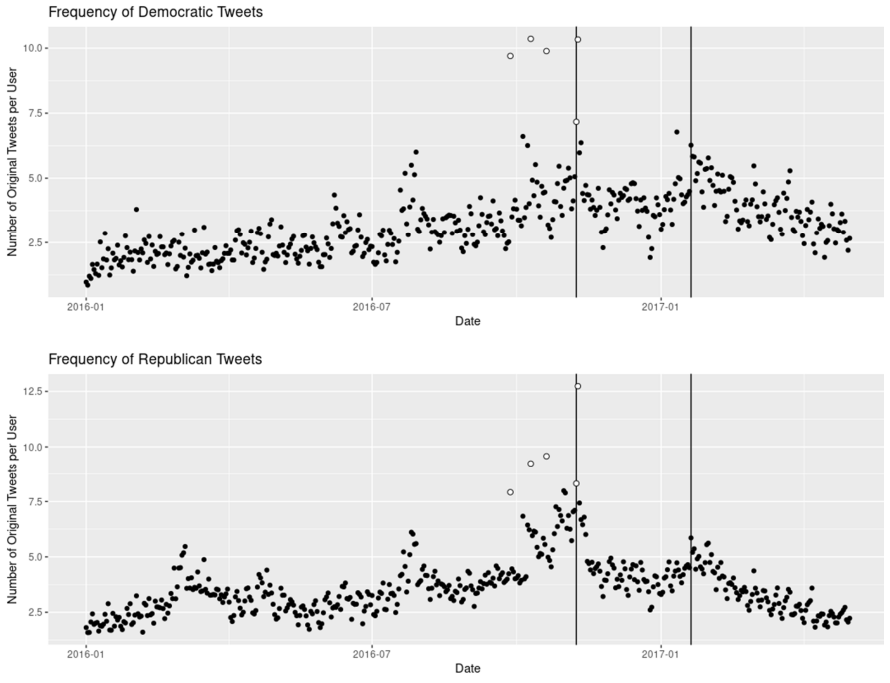


Figure 2 Average number of original tweets per day per Democrat (top) and Republican (bottom) users from 2016 through mid-2017. Vertical lines represent election day (November 8, 2016) and inauguration day (January 20, 2017). White points are the days with the highest frequency of tweets for Democrats and Republicans.

After observing fairly convincing evidence that our set of users are tweeting about political events, we next consider sentiment of original tweets, measuring how the users reacted to those events. We find that while frequency of tweets among our politically active users is mainly driven by political events, sentiment for both Democrats and Republicans is driven by both political and nonpolitical events. Large daily spikes in average sentiment for all tweets from Democrats and Republicans correspond to holidays, such as Christmas and Thanksgiving, and a large daily drop is likely in response to a mass shooting, as can be seen in Figure 3.

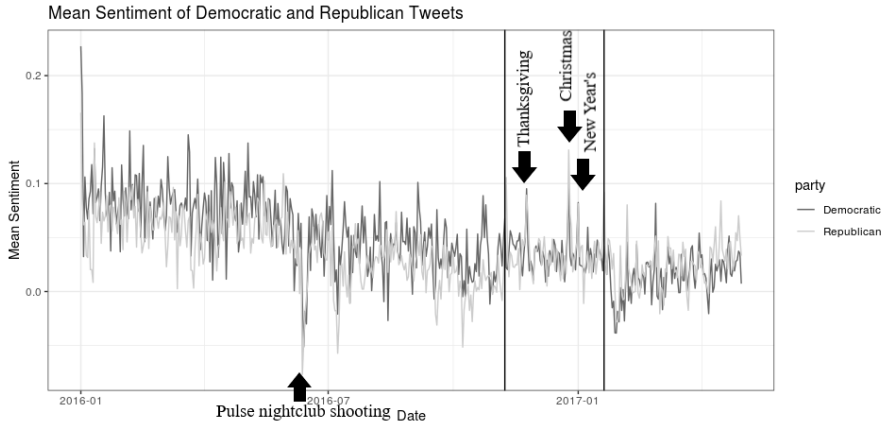


Figure 3 Average daily sentiment for Democrats (dark grey line) and Republicans (light grey line) from May 2016 through May 2017. Vertical lines represent election day (November 8, 2016) and inauguration day (January 20, 2017).

Many of the events that affect the sentiment of tweets of both our Democrats and Republicans occur outside of the political realm. Therefore, with the idea that Democrats and Republicans react to holidays and tragedies with similar sentiment, we are instead interested in the difference in sentiment between Democrats and Republicans. By taking the difference in sentiment, we conceivably remove "cultural noise" while enhancing the political signal. Figure 4 shows the daily difference in the mean sentiment of Democratic and Republican tweets from two months before the election through two months after the election. There is a clear drop the day after the election, and there appears to be an overall change when comparing difference in sentiment from before the election to after the election: Democrats are generally happier before and Republicans happier after. Presumably because the election results were a surprise for many, the notable change in difference in sentiment between Democrats and Republicans was immediate as opposed to gradual.

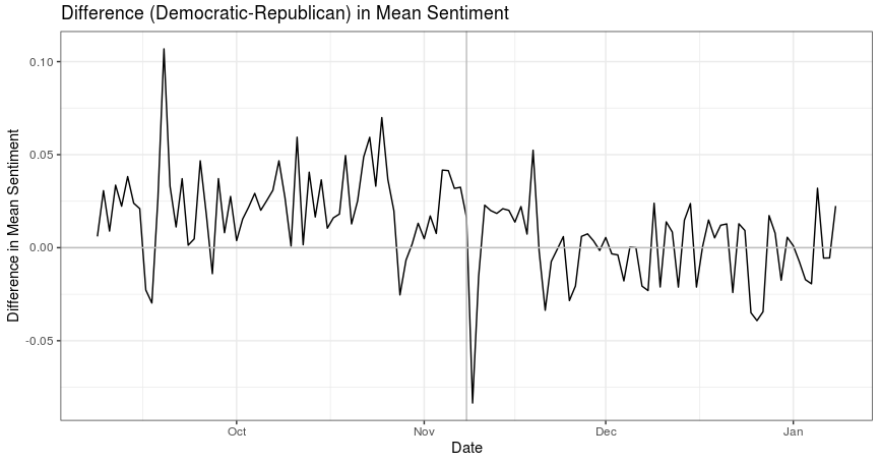


Figure 4 Difference in average sentiment between Democrats and Republicans (Democrats minus Republicans) from two months before the election (September 8, 2016) through two months after the election (January 8, 2017). The vertical line is election day (November 8, 2016).

While Figure 4 suggests a genuine difference in sentiment between our set of Democrats and Republicans from before the election compared to after the election, this change in sentiment is arguably relatively small. We look specifically at users who are vocal about politics and have fairly clear political party affiliation. We thought that the 2016 presidential election would be one of the most consequential events on Twitter for these users, and the observed difference in sentiment in Figure 4 is less pronounced than we might have imagined for such a set of users.

Discussion

If social media data is to be used to supplement or replace surveys tracking public opinion, there must be sufficient evidence that the social media data is indeed a valid way of measuring public opinion. This includes evidence that we are indeed tracking the signal of interest, a high signal to noise ratio, and stability of the relationship over time. We address these issues in accomplishing our two main goals: developing a framework to interpret an observed relationship between surveys of public opinion and tweets containing some keyword, and finding evidence of a political signal when following Twitter users longitudinally.

We found the correlation between sentiment of "Trump" tweets and presidential approval, 0.516, by optimizing smoothing of sentiment and lag between

survey responses and tweets. We developed a framework to interpret the strength of this observed correlation by comparing it to 497 placebo correlation obtained by performing the same analysis, but with tweets containing everyday words. The correlation of 0.516 was not especially strong in comparison with the reference distribution. This shows that there is a high level of noise in Twitter data; many of the placebo correlations, which should consist of nearly pure noise, were as high as the correlation between "Trump" tweets and presidential approval. As an alternative method to tracking tweets that contain the word "Trump", we proposed following politically active users longitudinally over time. We found evidence of a political signal when classifying users as Democrat or Republican based on the accounts they follow. When tracking the frequency of their tweets over time, we found a clear political signal, with frequency of tweets spiking at political events. The difference in sentiment between Democrats' and Republicans' tweets also changed immediately following the 2016 election. Noticeable changes in the tweeting patterns of our set of users around political events confirms that we are indeed capturing our political signal of interest. This is consistent with previous results that found events in Twitter data, for example frequency of "Obama" and "Romney" tweets leading up the 2012 presidential election (Barberá & Rivero, 2015) and sentiment of "Obama" tweets spiking on Obama's birthday (Pasek, McClain, Newport, & Marken, 2019). However, given that the election was what we assumed to be one of the clearest signals on Twitter for this particular set of users, the change in sentiment is relatively small. The conclusions of both the cross-sectional and longitudinal analyses are in agreement that finding strong, clear, long-term signals in sentiment of Twitter data is not a trivial task. We do, however, have evidence that Twitter does respond to the onset of events on a short time scale, such as spikes in sentiment around holidays and spikes in frequency around larger political events. Given the tentatively encouraging results from the longitudinal section, future analyses tracking an appropriate set of users over time may be more effective at recovering a continuous public opinion trend over time than tracking tweets containing a given word.

While we only considered social media data extracted from Twitter, similar methods can be applied to data extracted from other social media platforms. For example, we can interpret the relationship between Reddit posts containing the word "Trump" and presidential approval using our placebo analysis framework. Tracking social media users from other platforms over time may also be a valid and fruitful method of extracting posts to analyze. Additionally, classifying users into various categories based on what they follow on the social media platform (users, subreddits, etc.) can be an effective method of collecting an appropriate set of users to track.

Creating a post on social media is in many ways different from responding to a survey question (Schober, Pasek, Guggenheim, Lampe, & Conrad, 2016), involving different psychological processes, reasons for posting, and considerations of

the audience. As one example, the demographics of social media platforms do not reflect the demographics of the general population (Wojcik and Hughes, 2019); this non-probability aspect of Twitter may be one of the reasons why tracking long-term trends in public opinion has been so elusive (Salganik, 2019). All of these differences have the potential to introduce bias, and completely removing this bias from social media data is perhaps a nearly impossible task.

While we have found no evidence that tweets containing a given keyword reliably track public opinion, we still believe there is potential for social media data to be utilized for this purpose. The results of our longitudinal analysis suggest that there is a real, if weak, signal in Twitter data, and a future line of work could make use of that signal. This seems unlikely to replace traditional public opinion surveys, but could potentially supplement surveys. Smith and Gustafson provide an example of supplementing election polls with Wikipedia page views of candidates to more accurately predict election results (Smith & Gustafson, 2017). Many challenges lie ahead, but with the right methods, there is potential for social media data to improve upon traditional methods of capturing public opinion.

Data Availability

Presidential approval was downloaded from the website FiveThirtyEight, available at https://projects.fivethirtyeight.com/trump-approval-ratings/?ex_cid=rrpromo. Data and scripts for replicating all analyses in this paper can be found at https://github.com/robynferg/Tracking_Presidential_Approval_with_Twitter. The Twitter data available online used in the placebo analysis gives the daily average sentiment for tweets containing each of the placebo words. To protect the privacy of the politically active users, we have blinded the user name and tweet content in the data set available online.

Software Information

Sentiment calculations using Vader were performed in Python version 3.65. All other analyses were performed in R version 3.5.1.

References

Athey, S., & Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives, 31*(2), 3-32.

Barberá, P., & Rivero, G. (2015, December 1). Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review, 33*(6), 712-729.

Cody, E. M., Reagan, A. J., Sheridan Dodds, P., & Danforth, C. M. (2016, August 5). *Physics*. Retrieved from arXiv.org: https://arxiv.org/pdf/1608.02024.pdf

Cohen, R., & Ruths, D. (2013, June). Classifying political orientation on Twitter: It's not easy!. In Seventh international AAAI conference on weblogs and social media.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011, October). Predicting the political alignment of twitter users. In 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing (pp. 192-199). IEEE.

Conrad, F. G., Gagnon-Bartsch, J. A., Ferg, R. A., Schober, M. F., Pasek, J., & Hou, E. (2019). Social Media as an Alternative to Surveys About the Economy. *Social Science Computer Review*. doi:https://doi.org/10.1177/0894439319875692

Daas, P. J., & Puts, M. J. (2014, September). Social Media Sentiment and Consumer Confidence. *European Central Bank Statistics Paper Series*(5).

De Heer, W., & De Leeuw, E. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. Survey nonresponse, 41, 41-54.

De Vreese, C. H., Boukes, M., Schuck, A., Vliegenthart, R., Bos, L., & Lelkes, Y. (2017). Linking survey and media content data: Opportunities, considerations, and pitfalls. *Communication Methods and Measures*, *11*(4), 221-244.

DiNardo, J. E., & Pischke, J.-S. (1996). The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too? *NBER Working Paper Series*.

Golder, S. A., & Macy, M. W. (2011, September 30). Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science, 333*(6051), 1878-1881. doi:10.1126/science.1202775

Hsieh, Y. P., & Murphy, J. (2017). Total Twitter Error. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, . . . B. West, *Total Survey Error in Practice* (pp. 23-46). Hoboken, New Jersey: Wiley.

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. Ann Arbor, MI.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). Evaluating online nonprobability surveys. Pew Research Center. Available at: http://www.pewresearch. org/2016/05/02/evaluating-online-nonprobability-surveys/(accessed September 2016).

Murphy, J., Link, M. W., Hunter Childs, J., Langer Tesfaye, C., Dean, E., Stern, M., . . . Harwood, P. (2014). Social Media in Public Opinion Reserach: Executive Summary of the AAPOR Task Force on Emerging Technologies in Public Opinion Research. *Public Opinion Quarterly, 78*(4), 788-794.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the Fourth International AAAI Conference on WEblogs and Social Media*, 122-129.

Pasek, J., McClain, C. A., Newport, F., & Marken, S. (2019). Who's Tweeting About the President? What Big Survey Data Can Tell Us About Digital Traces? *Social Science Computer Review*.

Pennacchiotti, M., & Popescu, A. M. (2011, August). Democrats, republicans and starbucks afficionados: user classification in twitter. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 430-438).

Pasek, J., Yan, H. Y., Conrad, F. G., Newport, F., & Marken, S. (2018). The Stability of Economic Correlations Over Time: Identifying Conditions Under Which Survey Tracking Polls and Twitter Sentiment Yield Similar Conclusions. *Public Opinion Quarterly, 82*(3), 470-492.

Salganik, M. J. (2019). Bit by bit: Social research in the digital age. Princeton University Press.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016, January 13). Social Media Analyses for Social Measurement. *Public Opinion Quarterly, 80*(1), 180-211. doi:https://doi.org/10.1093/poq/nfv048

Silver, N. (2017, March 2). *How We're Tracking Donald Trump's Approval Ratings*. Retrieved from FiveThirtyEight: https://fivethirtyeight.com/features/how-were-tracking-donald-trumps-approval-ratings/

Smith, B. K., & Gustafson, A. (2017, May 06). Using Wikipedia to Predict Election Outcomes: Online Behavior as a Predictor of Voting. *Public Opinion Quarterly, 81*(3), 714-735. doi:https://doi.org/10.1093/poq/nfx007

Vijayaraghavan, P., Vosoughi, S., & Roy, D. (2017, July). Twitter demographic classification using deep multi-modal multi-task learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 478-483).

Wojcik, S., & Hughes, A. (2019). Sizing up Twitter users. Pew Research Center. Available at: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/(accessed June 2020).

Appendix A: Sentiment of "Trump" Tweets

The daily variation in mean sentiment of "Trump" tweets is intrinsic to the Twitter data itself; it is not due to the fact that we have sampled 1000 tweets per day. To demonstrate this, we plot the unsmoothed daily average sentiment for the first 100 days with associated error bars. That is, we plot the 95% confidence intervals for the population mean sentiment of all "Trump" tweets. This can be seen in Figure A1. We only plot the first 100 days to more easily see the change day-to-day. The confidence intervals for one day to the next fairly frequently do not intersect. While we only show the first 100 days, the pattern of non-overlapping confidence intervals continues throughout the entire time frame.

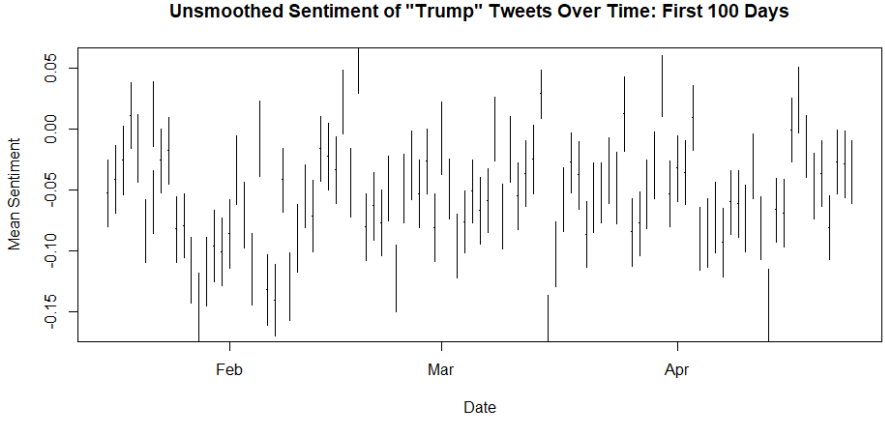


Figure A1 Daily sentiment of "Trump" tweets over time with associated confidence intervals.

Appendix B: Optimal Values and Changes Over Time

When finding the optimal correlation for the 497 placebo words, we obtain 497 optimal k and L values. Figure B1 shows the optimal smoothing and lag parameters for each of the placebo words. Many of the optimal smoothing parameters are at the maximum allowed by our smoothing window. This is a cautionary messages: too much smoothing can lead to artificially inflated correlations.

Throughout the time period of performing the analysis and writing this paper, we re-ran the analyses several times as newer data became available. Results often depend on the last data point available in the analysis. Consider finding the optimal correlation between sentiment of "Trump" tweets and presidential approval when the last data point available ranges from May 20, 2017 to August 25, 2019. For each of those end dates we find the smoothing and lag parameter that leads to the maximum absolute correlation. Figure B2 shows the maximum absolute correlation (thick line) and the correlation with 45 day smoothing and 30 day lag (dashed line) change over time. Figure B3 shows the optimal smoothing and lag values that produce the maximum absolute correlation as the end date of the data changes. The optimal smoothing and lag parameters stabilized around mid-2018.

The placebo words with correlations greater than our observed correlation of 0.516 are: "hell", "wanted", "retweet", "enough", "17", "000", "like", "name", "piece", "help", "ppl", "black", "room", "1st", "find", "story", "lie", "let", "twitter", "might", "talk", "together", and "walk". None of these placebo words are meaningfully related to presidential approval.

The reference distribution also changes as end date changes. Figure B4 shows how the proportion of placebo correlations that are more extreme than the correlation between sentiment of "Trump" tweets and presidential approval changes as the end date of the data changes. Around mid-2018, this proportion stabilizes to between 0.05 and 0.10. If we change the maximum lag to 7 days, we obtain similar results, see Figure B5.

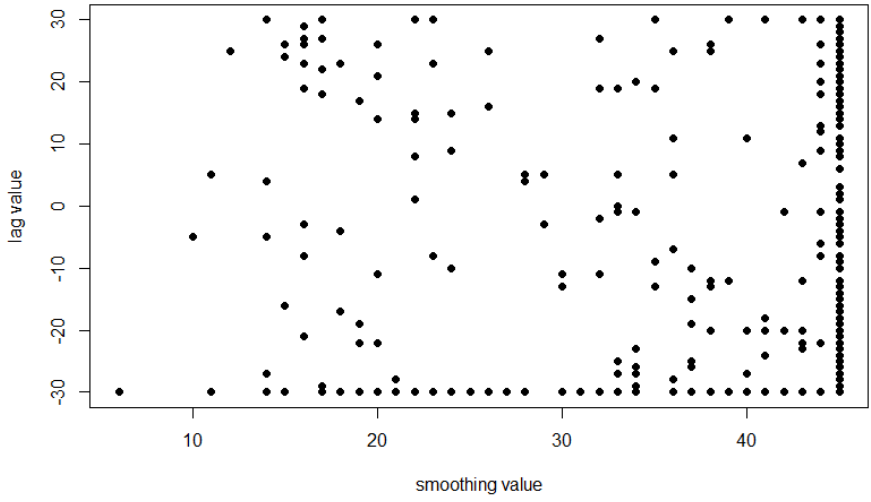


Figure B1 Locations of optimal smoothing and lag parameters between the 497 placebo words and presidential approval. Each point represents where the maximum correlation occurs for one of the 497 placebo words appearing in the Twitter corpus every day.

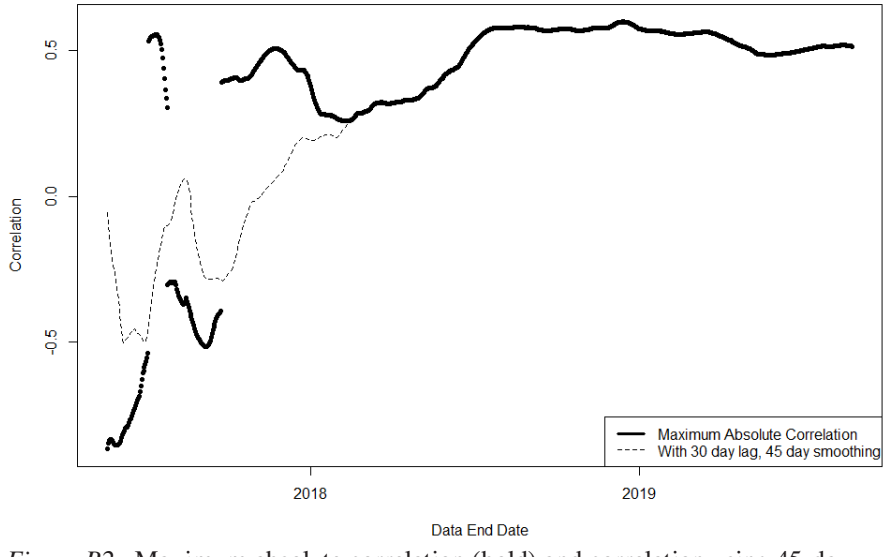


Figure B2 Maximum absolute correlation (bold) and correlation using 45-day smoothing and 30-day lag (dashed) as end date of data changes.

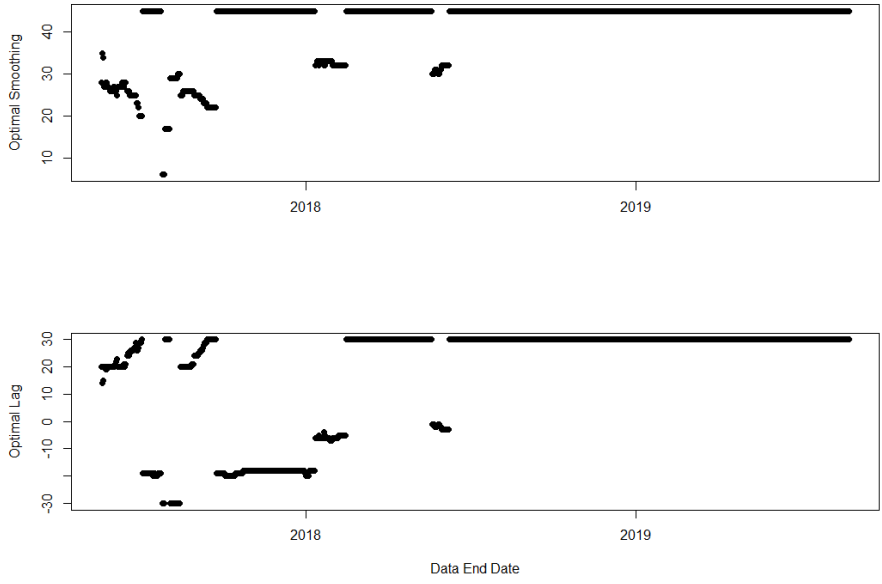


Figure B3 Optimal smoothing (top) and lag (bottom) parameters as end date of data changes.

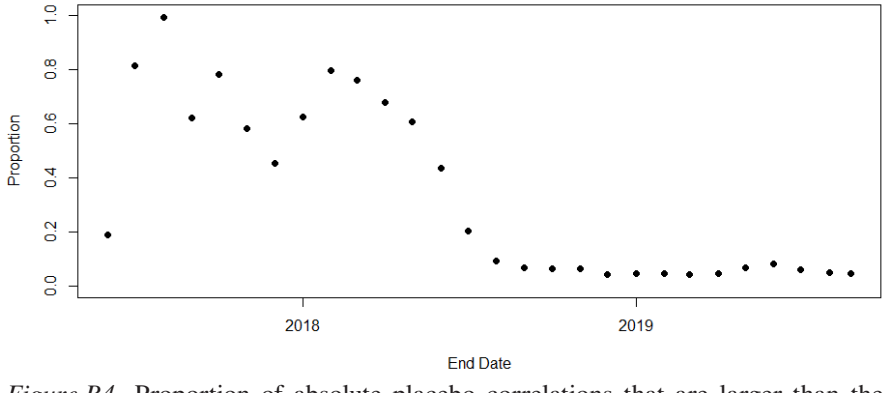


Figure B4 Proportion of absolute placebo correlations that are larger than the correlation between "Trump" tweets and presidential approval as end date of data changes, from June 1, 2017 to August 25, 2019.

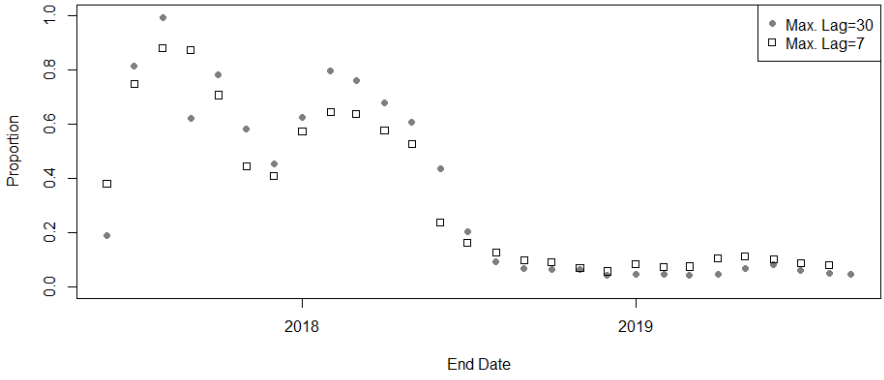


Figure B5 Proportion of absolute placebo correlations that are larger than the correlation between "Trump" tweets and presidential approval as end date of data changes, from June 1, 2017 to August 25, 2019, when maximum lag is 7 days compared to 30 days. Changing lag windows does not drastically change our interpretation of the strength of correlation between sentiment of "Trump" tweets and presidential approval.

Appendix C: Identifying Politically Active Users and Political Beliefs

The set of politically active users was created using a corpus of tweets provided to us by Sysomos. All tweets in this corpus contained the word "jobs" and were used in a previous analysis unrelated to this paper (see Conrad et. al. (2019)). We created an algorithm to classify "jobs" tweets into various categories, one of which was 'news/politics', based on the words within a tweet. See Conrad et. al. (2019) online appendix for details on this algorithm. We take a random sample of size 15,000 of the users whose "jobs" tweet was classified as political and retrieved their 2016 tweets history. If a user produced at least 20 original (non-retweets) in 2016, at least 10 of which contained a political word, we consider that user a 'politically active user'. While this method of classifying tweets as political or not surely mislabeled true political tweets as non-political, we have a high level of certainty that the tweets classified as political were truly political.

By looking at many self-provided profile descriptions, we created a list of commonly found words that make the user's political party known: "conservative", "Trump", "MAGA", "NRA", "constitution", "Republican", "Libertarian", "Democrat", "liberal", "Hillary", "Clinton", "Obama", "progress*", "Bern*", "resist*", "president". If a politically active user's self-provided profile description contained one of these words, we hand-classify that user as belonging to one of the two major political parties in the US: Democratic or Republican. These users were explicitly clear in their profile description about their political beliefs or about which candidate they did or did not support in the 2016 presidential election. We classify self-described libertarians as Republicans, and classify self-described socialists as Democrats. We classify Never-Trump Republicans as Republicans, and classify Never-Hillary Democrats as Democrats. This gives our training set of 170 Democrats and 393 Republicans.

We use a random forest as the classifier, with the covariates being accounts that at least 30 of the politically active users with known political affiliation follow. We give the confusion matrix of the random forest and the variable importance plot. Table C1 contains the confusion matrix; only 9% of the Democrats were incorrectly classified as Republicans by the random forest, and only 0.85% of the Republicans were incorrectly classified as Democrats. Figure C1 gives the variable importance plot of the random forest classifier. Out of the top 30 accounts shown in the variable importance plot, all are in some way political, either politicians, family members of politicians, or political commentators.

The set of politically active users was created in mid-2017. Twitter has since deleted many bot accounts that had the goal of influencing other users' political

opinions. We want to ensure that we have not gathered multiple bot accounts in our set of politically active users; we want the opinions of real people.

Out of the 1485 politically active users identified in mid-2017, 99 accounts were unable to be scraped in May 2018. These are split fairly evenly across Democrats and Republicans: 7% of Republicans' and 5% of Democrats' tweets were not able to be gathered using the Twitter API in May 2018. However, this does not mean the account was a bot; users can choose to delete their account at any time, can make their account private, or have their account suspended by Twitter, all of which would result in the account being inaccessible using the Twitter API.

NBC published a list of 453 bot users and tweets from those bots (Popken, 2018). Our list of Democrats and Republicans did not contain any of these known bots.

References

Popken, B. (2018, February 14). *Twitter deleted 200,000 Russian troll tweets. Read them here.* Retrieved from NBC: https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

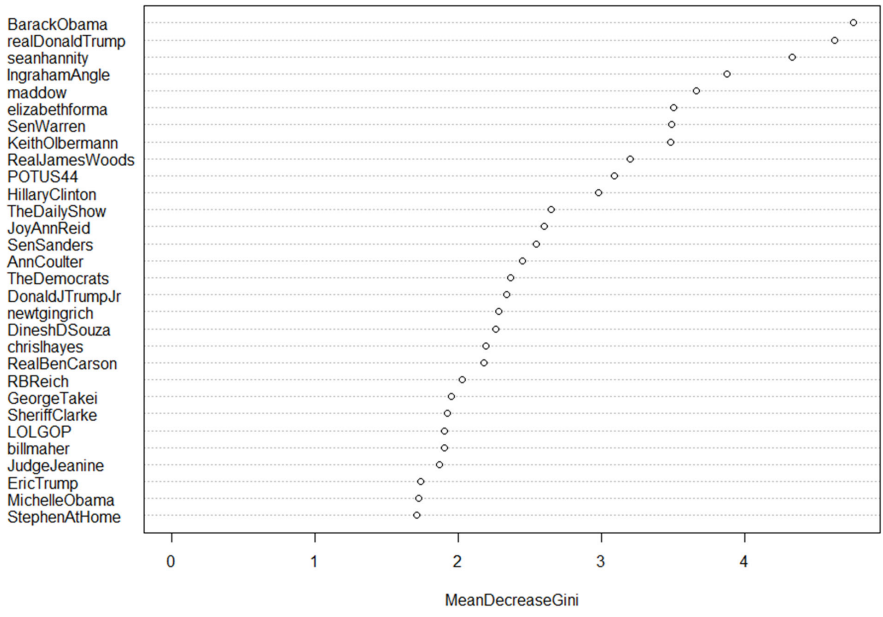


Figure C1 Variable importance plot of Twitter accounts used in classifying users as Democrat or Republican. All of the top 30 accounts above used to classify are political, most being either politicians (e.g. BarackObama, realDonaldTrump), political commentators (e.g. seanhannity, IngrahamAngle, maddow), or family members of politicians (e.g. DonaldJTrumpJr, MichelleObama).

Table C1 Random forest confusion matrix. Actual party affiliation corresponding to the hand classification; predicted party affiliation corresponding to the random forest out-of-bag prediction.

		Predicted		
		Democrat	Republican	Classification Error
Actual	Democrat	160	10	0.090
	Republican	5	388	0.0085

Appendix D: Changes in Positive and Negative Sentiment over Time

To get a more detailed understanding of what was driving the change in difference in sentiment, we looked at how the positive and negative sentiments changed over time. When looking at the difference in means of the positive tweets, there is a clear drop immediately following the election, and a smaller drop around the inauguration. However, no such change is seen in the difference in negative tweets (see Figure D1). The overall change in difference in sentiment was driven by Republicans' positive tweets becoming more positive post-election.

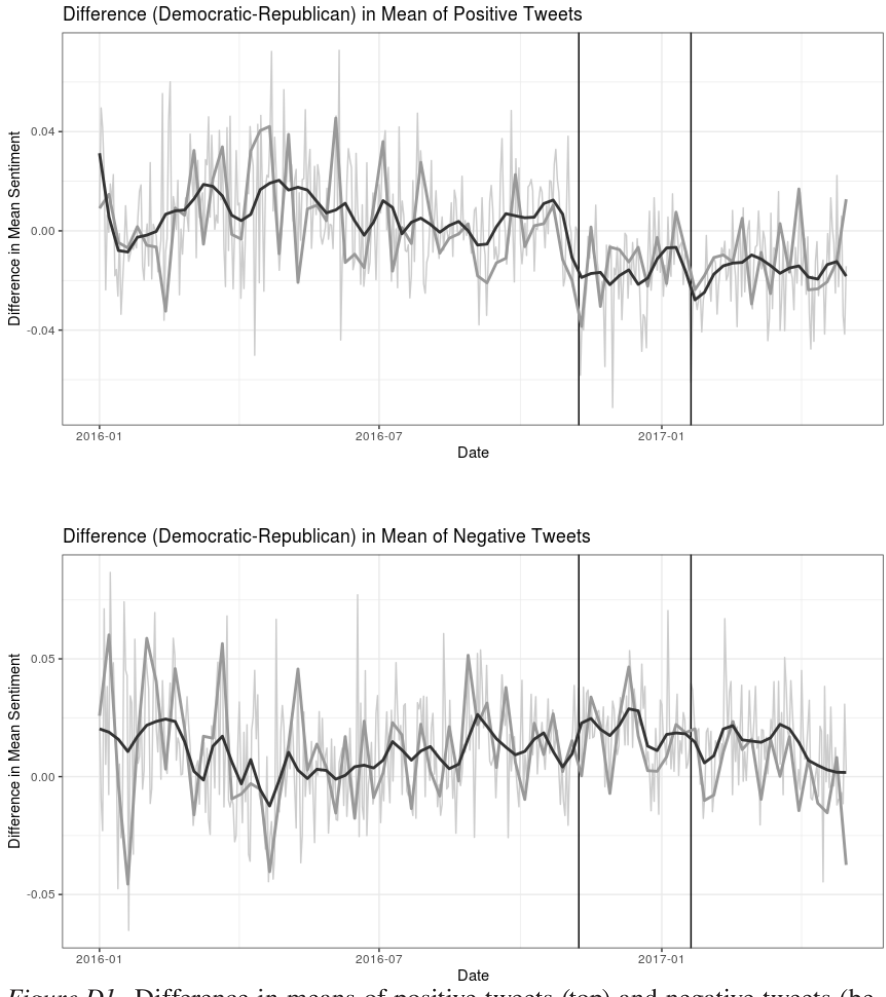


Figure D1 Difference in means of positive tweets (top) and negative tweets (below) for Democrats minus Republicans. The vertical lines are election day (November 8, 2016) and inauguration day (January 20, 2017). The different shaded lines are for various smoothing levels to more easily see how sentiment changes over time. The notable change in positive difference (top) post-election is due to Republicans' positive tweets became more positive post-election.

methods, data, analyses | 2021, Vol. 15(2), pp. 241-264 DOI: 10.12758/mda.2021.03

A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model

Ralf Münnich[1], *Rainer Schnell*[2], *Hanna Brenzel*[3], *Hanna Dieckmann*[1], *Sebastian Dräger*[1], *Jana Emmenegger*[3], *Philip Höcker*[2], *Johannes Kopp*[1], *Hariolf Merkle*[1], *Kristina Neufang*[1], *Monika Obersneider*[2], *Julian Reinhold*[2], *Jannik Schaller*[3], *Simon Schmaus*[1], *& Petra Stein*[2]

[1] *Trier University, Germany*

[2] *University of Duisburg-Essen, Germany*

[3] *Federal Statistical Office, Germany*

Abstract

Microsimulation models are widely used to evaluate the potential effects of different policies on social indicators. Most microsimulation models in use operate on a national level, disregarding regional variations. We describe the construction of a national microsimulation model for Germany, accounting for local variations in each of the more than 10,000 communities in Germany. The database used and the mechanisms implementing the population dynamics are described. Finally, the further development of the database and microsimulation programs are outlined, which will contribute towards a research lab that will be made available to the wider scientific community.

Keywords: microsimulation methods, spatial microsimulation, social simulation, multivariate modelling, multi-source modelling, synthetic data generation

What are the effects of changes in the demographic profile of the population on family formation processes? How does tax legislation impact on tax revenues? Do changes in women's employment alter the way in which the elderly are cared for? How do different qualifications and work biographies influence pensions? How will integration patterns develop in the future in the context of demographic transition regarding upcoming generations of migrants? These are some examples of problems studied using microsimulation techniques in past decades (Stein & Bekalarczyk, 2016; O'Donoghue & Dekkers, 2018; Schnell & Handke, 2020; Zwick & Emmenegger, 2020).

However, due to limited data availability and high computational complexity, such simulations were mostly done at national levels of analysis. Given the fact that social and economic change is often different for some regions and subgroups of the population, the demand for detailed modeling is increasing. Therefore, highly detailed regional datasets covering the whole population are needed. A prime example is schooling: The demand for elementary schools varies at the local level. For secondary schools, increasing population heterogeneity requires complex school planning which accounts for the diversity of schooling demands. A second example is care for the elderly, where information regarding the distance between parents and their children, who are potential care providers, is essential for modeling the demand for care.

In the Netherlands or the Scandinavian countries, population covering register datasets could provide data for regional microsimulations. In Germany, however, hardly any datasets are available that foster regional microsimulation modeling. Either the sample size is too small for regional models or data protection regulations do not allow the use of low-level identifiers in population covering datasets. Therefore, modeling tasks requiring low-level information are challenging. The MikroSim model described in this paper is aiming for a dynamic microsimulation of Germany down to each municipality.[1] The model is based on a highly detailed dataset build from many different sources, such as survey, administrative, and other data. Therefore, a series of different methods of data-integration and small area

1 The corresponding MikroSim-project, in which this model was developed and that is funded by the German research association, is described in detail in Münnich, Schnell, et al., 2020.

Acknowledgements

The research unit FOR 2559 MikroSim is funded by the German Research Foundation. The principal investigators are Ralf Münnich (Speaker), Rainer Schnell (Co-Speaker), Johannes Kopp and Petra Stein. The research unit cooperates with the German Federal Statistical Office, with Hanna Brenzel as primary contact, succeeding Markus Zwick.

Direct correspondence to

Ralf Münnich, Trier University
E-mail: muennich@uni-trier.de

estimation have been used for building the dataset. Modeling the dynamic processes is achieved by estimating transition probabilities using different statistical methods. The MikroSim model consists of different modules, simulating births, deaths, marriage, education, and other dynamic processes.

The structure of the present paper is as follows: We begin with a short overview on the history and the different approaches to microsimulations in general in section *Microsimulation Modeling*. Subsequently, we present *The MikroSim Model*. The subchapters of this section include the generation of the synthetic dataset of the German population as the base dataset of the model, the construction principles, the sequence in which the simulation modules are ordered and an overview on the specific conceptualization of each module. Some examples of modules that are to be implemented are described in section *Application Modules*, as well as a short examples on the questions that can be answered using microsimulation methods. A summary and an outlook on future developments concludes the paper.

Microsimulation Modeling

The beginning of microsimulation in economics and social sciences dates back to the 1950s when Guy H. Orcutt published the paper "A New Type of Socio-Economic System" (Orcutt, 1957). He criticizes the limited usefulness of macrosimulations due to the focus on aggregates and the inability to consider nonlinearities and discontinuities in individual behavior. He advocates a new type of modeling that focuses directly on micro-units, such as individuals, households, and firms. *This new type of model consists of various sorts of interacting units which receive inputs and generate outputs. The outputs of each unit are, in part, functionally related to prior events and, in part, are the result of a series of random drawings from discrete probability distributions* (Orcutt, 1957).

Thus, the main focus of microsimulation is to look at the smallest unit of a system. Li and O'Donoghue (2013) describe microsimulations as *a tool to generate synthetic micro-unit based data, which can then be used to answer many "what-if" questions that, otherwise, cannot be answered.* These questions are usually understood as the investigation of different scenarios, such as different social and tax systems and behavioral assumptions. Contrary to macro simulations, not only single target values but complex interrelations and distributions within the system can be investigated. According to Li and O'Donoghue (2013), microsimulations can be split into two tasks. The first step is to generate a high-quality dataset with the relevant variables of interest in the necessary geographic depth. In a second step, a set of scenarios is performed on this dataset in order to answer the what-if questions.

The basis of any microsimulation is a dataset – the so-called base population – that contains micro-level information about the system of interest. Socioeconomic questions usually require information on individuals and households. Due to the easy availability and large amount of information, survey datasets are mainly used as base populations (Li & O'Donoghue, 2013; Burgard, Dieckmann et al., 2020). However, this kind of data contains only a relatively small number of individuals and allows only very limited regionalized analyses. Larger datasets from administrative sources and census data often contain a limited set of variables. Hence, (partially) synthetic base populations have been increasingly used for regionalized models recently. The methods used to create small-scale datasets are often described as small-area or spatial microsimulation techniques (Tanton, 2014; Rahman & Harding, 2016).

Another distinction in microsimulations relates to the temporal component. In static microsimulations, there are usually no changes in individual states during the course of the simulation. The immediate distributional impact of (political) changes is evaluated without reference to the time dimension. In this case, it is assumed that the characteristics of the population of interest do not change rapidly (Merz, 1991). Thus, this kind of modeling is primarily suitable for short- and medium-range predictions. To implement a temporal component, re-weighting and uprating/deflating techniques can be implemented. In a re-weighting process, the survey weights are calibrated to exogenously given aggregate data of another time period while uprating/deflating changes the specific variables (for example specific income components) directly (Merz, 1991; Sutherland, 2018).

In dynamic models, micro-units interact and evolve over a temporal horizon. This type of simulation can account for micro-level dependencies and complex interaction allowing long-term projections and time-dependent behavior simulations. The focus is on sophisticated ceteris paribus analyzes over time under an approximation to real-world complexity. The so-called ageing process can either be continuous or discrete. In discrete time models, the base population is aged considering discrete – mainly annual or monthly – time intervals and events are realized in each period using transition probabilities. Continuous microsimulations, on the other hand, allow events to occur at any point in time until the simulation horizon is reached. Instead of transition probabilities, the simulation is usually based on survival analysis (Li & O'Donoghue, 2013; Burgard, Dieckmann et al. 2020).

For a more detailed methodological differentiation of dynamic microsimulation models, we refer to Li and O'Donoghue (2013) and Hannappel and Kopp (2020).

The focus in this paper is on discrete time dynamic microsimulations. The simulated events can either be deterministic or stochastic. Deterministic changes of states are, for example, the ageing of individuals in each period or the loss of income after the termination of employment. However, dynamic microsimulations

are usually characterized by the fact that all events depend, directly or indirectly, on one or more stochastic processes. The simplest way to simulate changes is based on first-order Markov processes, where the occurrence of an event depends exclusively on the state of the previous period. The probabilities are organized in transition matrices, which are usually differentiated according to socio-demographic and socio-economic characteristics. Transition matrices can be easily estimated using conditional distributions. However, the most common way to obtain individual transition probabilities is to estimate logit (multinomial) regression models based on panel data. In the models, the dependent variable can either be conditioned to the state of the previous period or the lagged variable can be included directly. The individual transition probabilities are predicted in the simulation process using the estimated model parameters. The simulation of state changes within the simulation is usually organized in modules. In each period, all individuals run through each module in a fixed order. Within each module, specific events are simulated. A module can be understood as a function which uses the base population as input and returns the updated population (Burgard, Krause, Merkle et al., 2019). A more detailed explanation of the estimation process of transition probabilities and the simulation process can be found in Burgard, Krause, Merkle et al. (2019) and Burgard, Krause, and Schmaus (2020).

The MikroSim Model

Base Data

A synthetic dataset is used as the base dataset for the microsimulation model MikroSim. In general, the purpose of synthetic datasets is to mimic a non-accessible or non-existing dataset so that the relevant characteristics of a synthetic dataset matches the characteristics of the underlying population as close as possible (cf. Münnich & Schürle, 2003; Münnich, Gabler et al., 2012; Kolb, 2013; Alfons, Kraft et al., 2011). The characteristics to be matched are distributional parameters, correlations, cluster effects, and totals.

The generated dataset is based on an anonymized national register of residents that has been used for methodological research for the census 2011. The German population with respect to all 11,339 municipalities is modeled.[2]

Since the register of residents contains only a few variables, additional data is generated. Since most data stems from German official statistics, data evaluation methods are performed within the statistical office to ensure confidentiality and privacy or, alternatively, using scientific use files. No record linkage between

2 A previous project (REMIKIS) modeled the region Trier using a similar modeling strategy (Burgard, Krause, Merkle et al., 2019).

official microdata and the synthetic dataset or microdata from other sources takes place. The generation of additional variables is formally defined recursively as

$$f(x, y, z) = f(x) \cdot f(y|x) \cdot f(z|x, y)$$

The variable set x contains the basic demographic variables such as age, gender and marital status. Further sets of variables are included in the German Microcensus including x to provide the information on the conditional distribution $f(y|x)$. A second block of variables y comprises variables related to education and activity status. Further variable sets z include variables of special interest for MikroSim, such as care or migration related topics. The conditional distributions are, in general, modeled using the multinomial logit models (Alfons, Filzmoser et al., 2011; Kolb, 2013). For the data synthesis, cluster effects resulting from a positive correlation between household members are considered. The variables for the household members are generated considering a household type variable to account for these cluster effects.

Finally, the synthetic population for each municipality is adjusted to published census totals. Using simulated annealing (cf. Laarhoven & Aarts, 1987; Huang & Williamson, 2001; Williamson, 2012; Tanton, 2014), households are selected randomly and entered or deleted sequentially to minimize the differences between the synthetic population and the census totals.

Construction Principles and Module Ordering

The MikroSim model is designed as a closed population simulation model. Therefore, modules simulating the paths for individuals entering or leaving is key to obtaining realistic projections of the population. The modules providing these paths are mainly the modules Mortality, Births, and Regional Mobility, which are later described in more detail. The simulation model is implemented in R.

Most of the modules in the MikroSim model are based on statistical models to estimate individual transition probabilities for the micro-units. Mainly used to determine these individual transition probabilities are for example data tables as well as multinominal and binary logit regression models (Burgard, Krause, Merkle et al., 2020). Regional differentiations as well as rural-urban disparities are modeled using adequate auxiliary variables within the models.

The characteristics of the simulated population are updated once for each simulated year. Therefore, no information about the exact time of occurrence of an event within the simulated period is available. However, the occurrence of one event might determine other events (for example, a death triggers further changes). There are different strategies available to deal with such dependencies (van Imhoff & Post, 1998). In MikoSim, probabilities for many events are only estimated for those persons who are eligible for a change of state. The eligibility is modeled by

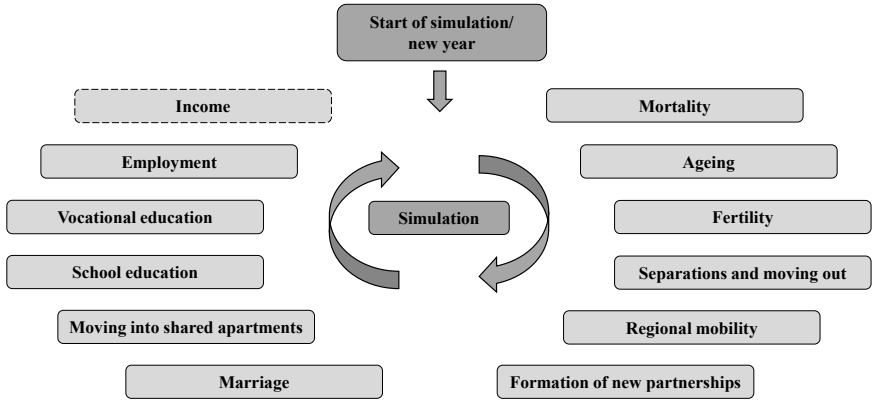


Figure 1 MikroSim Module Sequence

the ordering of the modules. For example, as the event of birth is simulated prior to the event of a marriage, the probability of marriage can be conditioned on the event of a birth (Burgard, Krause, Merkle et al., 2020). It is also possible to start with the marriage event and use this information for the prediction of the birth. From a theoretical perspective, not the order of modules but the modeling strategy is crucial for the simulation. Let $f(x, y)$ be the conditional distribution of the events birth $f(x)$ and marriages $f(y)$. There are two possibilities to reach the joint distribution:

$$f(x, y) = f(x) \cdot f(y|x) = f(y) \cdot f(x|y).$$

Nevertheless, it is always possible to take states from the previous time period into account in the estimation process. Regarding the example above, this means that the marital status in the current period can influence the transition probabilities for the birth of a child in the next period.

The sequence in which the modules are processed is shown in Figure 1.

The simulation of state transitions is conducted using random draws from the predictive distribution of the variables of interest (inversion method). First, the cumulative individual transition probabilities are calculated. Then, a uniformly distributed random number is drawn for each individual and the state is set to the value in which interval the random number lies. For example, let the transition probabilities for a full-time employed person to remain in this state be 0.70, to change into a part-time employment 0.2 and to get unemployed 0.1. The cumulative probabilities are $\{0.7, 0.9, 1\}$ and the random number is 0.83. Consequently, the person changes to a part-time job as $0.83 \in [0.7, 0.9)$.

One major challenge in dynamic microsimulation is in the fact that the transition probabilities are mainly estimated using sample data. Because of data limitations, the number of estimated events often does not match known benchmarks. In addition to the small number of observations, regional differentiations often cannot be made for data protection reasons. This problem is common in the field of microsimulation modeling since *no country has the ideal dataset for [...] estimating the parameters of all the processes in a dynamic microsimulation model* (Bækgaard, 2002). However, the application of small area methods are applied to provide accurate regional benchmarks (Rao & Molina, 2015; Münnich, Burgard, & Vogt, 2013).

To harmonize the individual transition probabilities with the known benchmark values on a macro level, alignment methods are applicable. In the context of dynamic microsimulations, various methods to adjust the transition probabilities or the number of transitions are available (Bækgaard, 2002; Li & O'Donoghue, 2014; Klevmarken, 2008; Stephensen, 2016; Burgard, Krause, & Schmaus, 2020). However, the methods differ considerably with regard to their applicability and functionality. A simple and well performing method is logit-scaling, where the transition probabilities are calibrated to a benchmark using a bi-proportional scaling algorithm. The solution corresponds in a logit framework to the adjustment of the intercept and leads to a solution which minimizes the Kullback-Leibler divergence between the estimated and calibrated probabilities (Stephensen, 2016). Klevmarken (2008) suggests a method to align the parameters by minimizing the quadratic difference between the estimated and adjusted values weighted by the inverse variance-covariance matrix. A constraint likelihood approach where the parameters are aligned while maximizing the original likelihood function aslo shows very good results (Burgard, Krause, & Schmaus, 2020). In the first version of the simulation, alignment is conducted using logit-scaling. Currently, other methods are also being implemented and can be applied via function arguments.

The first module in the MikroSim model simulates widowhood for married persons not living in the same household with their spouse. Widowhood directly influences the possibility of the respective persons to enter the trailing modules such as separations or the formation of new partnerships. Updating this relationship status at the beginning can prevent an underestimation of widowhood within separated couples.

The Mortality module is placed prior to the Aging module since also newborns face a non-zero mortality risk. The event of death depends exclusively on age and gender. Following the Aging module, births are simulated as the first way to add new individuals to the population. The position of the Birth module at this early stage of the simulation is required since birth decisions must precede the birth event. Thus, birth probabilities are estimated mainly based on the characteristics in the previous period.

The Leaving Household module then simulates (1) relocations of adults from parental households, (2) shared apartments, and (3) dissolutions of households after separations. Since these relocation events often have a direct impact on regional mobility, migration across district borders is simulated subsequently.

New households are formed by creating new partnerships and shared apartments. Thus, persons can directly form new households within a simulation period after leaving a household or immigrating. Formal changes in the relationship status (divorces and marriages) are simulated separately.

Changes in school education, vocational education, and employment status are the final modules in the simulation process. An income module will be integrated soon.

The modules are based on a variety of different models, modeling methods, and datasets. Table 1 gives a brief overview, including the choice of independent variables.[3] The following sections describe the modules in more detail.

Modules

Mortality

The Mortality module is the first step in the simulation process. The probabilities for death are assigned according to sex and age of the simulated person using the life tables published by the Federal Statistical Office of Germany (Statistisches Bundesamt, 2020a). The event of death does not only affect the size of the population but also individual and household characteristics. For instance, a partner's family status has to be updated if the husband or wife dies and underage orphans living alone after a parent dies must be assigned to new households. In addition, widowhood for married persons who do not live in the same household cannot be updated deterministically and therefore has to be simulated. This model is based on the German Microcensus (Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, 2018a; 2018b; 2018c).

After removing the deceased individuals from the simulation and updating the family status, the age of all remaining units is increased by one year.

Fertility

The Fertility module in MikroSim simulates births in two steps. In the first step, for all women of fertile age (15–49 years) a probability of giving birth is estimated. The model uses individual characteristics of the women as well as characteristics of

3 A detailed explanation of the mechanisms sketched in Table 1 will be the subject of a different publication.

Table 1 Overview of the MikroSim Modeling approaches

Modules	Method	Y	X
Mortality	RF	Death	age; sex
	LR	Widowhood (not in cohabitation	age; age²; sex; working
	D	Widowhood (in cohabitation)	death of the partner
Ageing	D	Age	$age_{t+1} = age_t + 1$
Fertility	LR	Birth (woman between 15 and 49 years)	age; age²; school education; vocational education; working; working (partner); marriage; age of children in household; Eastern Germany
	LR	Twin birth	age
Seperations and Moving out	LR	Leaving cohabitation/marriage (up to 60 years of age)	age; age difference; age of children in household; marriage; school education (both partners); vocational education (both partners); citizenship homogamy; working; Eastern Germany
	LR	Leaving parental home (18 years and older)	age; sex; age² sex; sex; school education; vocational education; marriage; citizenship; Eastern Germany
	LR	Leaving a shared apartment (no relationship within household; from 18 to 35 years)	age; sex; age² sex; sex; school education; vocational education; marriage; citizenship; Eastern Germany
Regional Mobility (between districts)	RF	Outflow	age; sex; marriage
	RF	Inflow	age; sex; marriage
Formation of New Partnerships	LR	Cohabitation (separate models for men and women living alone; from 18 years)	children in household; children under 6 in household; age; age²; separated; widowed; school education; vocational education; working; Eastern Germany
	M	Finding a partner	matching on age, education, citizenship

Modules	Method	Y	X
Marriage	LR	Marriage (not married cohabitations)	age; age² (women); family status (both partners); ISCED-Level (both partners); working (both partners); children under 1 in household; Eastern Germany
Moving into shared Apartments	LR	Moving into a shared apartment (one-person household; from 18 to 35 years)	age; age²; sex
School Education	RF	Age of enrollment (5-year-olds)	federal states; sex
	MR	No qualification; lower secondary; secondary; technical college level; university entrance level	age; sex; school education (parents); vocational education (parents); working (parents); citizenship (parents); single-parent household
Vocational Education	MR	Enrolled; no vocational training; vocational training; tertiary	working (parents); vocational education (parents)
Employment	MR	Working; unemployed; inactive (from 15 to 65 years)	age; age²; sex; school education; vocational education; mar riage; citizenship; federal states; $working_{(t\ 1)}$; children in household; children under 3 in household
	LR	Full-time employment	age; age²; sex; school education; vocational education; marriage; citizenship; federal states; $working_{(t\ 1)}$; children in household; children under 3 in household;

RF	Relative Frequencies	OS	Official Statistics
LR	Logistic Regression	GM	German Microcensus
D	Deterministic	SOEP	German Socio-Economic Panel
M	Matching	AIDA	Growing up in Germany (Aufwachsen in Deutschland: Alltagswelten, AID:A)
MR	Multinomial Regression	NEPS	National Education Panel Study

other people living in the household, such as the employment status of a potentially existing partner or the age of the youngest child (see Table 1).[4]

In the second step, conditional on the event of a birth, twin births are simulated.[5] The estimation model includes only the age of the women. The sex of a simulated child is assigned in accordance with the known sex distribution of newborns. All other variables of a simulated person are initialized to reasonable values (such as age to zero and school or vocational qualifications to missing values).

Since the model is based on sample data lacking regional details, differences between observed and simulated birth rates may result for some districts. Therefore, the model for birth is calibrated to known birth rates of the German districts up to the last available data (Statistisches Bundesamt, 2019a).

Separations and Moving Out

In the MikroSim model, the events of individuals leaving households are simulated by three mechanisms: (1) separating from a partner, (2) leaving of the parental household and (3) moving out of a shared apartment. Modeling these three transitions using survey data is quite challenging, as individuals are either not tracked over time at all (e.g., in the German Microcensus) or only for subgroups (e.g., in the German Socio-Economic Panel, Goebel et al., 2019). Since the person moving out is usually only observed before a change of residential status occurs, identifying the cause of a departure is difficult using available data.

The first mechanism simulates the separation from a partner as ending a cohabitation. Therefore, only persons living in a partnership with cohabitation are considered. The probabilities are estimated with a logit model based on longitudinal Microcensus data for the years 2012 to 2014 using information of the partners and their partnership (such as age difference, for details see Table 1). In the case of a separation, new households are formed. These new households are initially single or multiple-person households if children are present. Currently, children are assigned to the mother, but future versions of the simulation will include predictive models for assigning children to new households.

The second mechanism models leaving the parental home. Only persons who are at least 18 years of age and still living in the parental household are considered. The probabilities are estimated with a logit model based on the same data as the mechanism described above. As predictors, age, the current level of education or vocational training and the relationship status are used (cf. Table 1). After leaving their parental homes, the individuals initially form new single-person households.

4 We plan to extend this model to include more independent variables such as nationality and regional information.

5 The birth of triplets or more children is not simulated due to the small number of cases.

The third mechanism models moving out of shared apartments.[6] Probabilities are estimated using a similar model and the same data as for the other two mechanisms. The number of people moving out is used for modeling the number of persons moving into shared apartments in later iterations of the simulation.

To prevent overestimating the number of single person households, the moved-out persons can form new households by entering either new partnerships or shared apartments.

Regional Mobility

The module Regional Mobility simulates between municipalities. Individuals leave the simulation population when moving out of the district and are added from a copy of the base population when moving into the district. This process is based on statistics produced by the Federal Statistical Office (Destatis, 2020). These statistics contain information on regional mobility broken down by age, sex, and relationship status. Subsequently, the probabilities for regional mobility are adjusted to known margins for each district level.

To prevent minors forming a household, regional mobility is additionally considered at the household level. Probabilities are estimated by Iterative Proportional Updating (Ye et al., 2009). Iteratively, the probabilities are adjusted to the frequencies of socio-demographic characteristics at the individual level in a randomized order and then scaled to a probability between zero and one (Stephensen, 2016). This is done using the base population to assign probabilities for leaving the district and on a copy of the base population to assign probabilities for moving into the district. Households that move into a district are selected from the copy of the population, which represents the remaining part of Germany. Newly arrived households are added to the base population. Outmoving households are removed from the population.

Formation of New Partnerships

The nuclear family (cohabitation of a mother, a father, and children) is still by far the most common family type in Western Europe. However, partners within the nuclear family increasingly remain unmarried (Schneider, 2015). Therefore, the simulation requires a module simulating the formation of new partnerships independent from the official status of a relationship (which is modeled in the Marriage model).

---

6 Shared apartments are defined by us as households containing at least two people aged between 18–35 without children or partnerships.

A module for the creation of new partnerships performs two simulation tasks: The entry into the partner market and the matching of new couples considering covariates.[7] There are different modeling approaches in other simulation models (cf. Perese, 2002; Zinn, 2012). MikroSim uses stochastic matching, therefore allowing less favorable partner combinations (for example, in large age differences). The module consists of two-steps. In the first step, for persons older 18 years not living in a partnership, a model based on German Socio-Economic Panel data (Goebel et al., 2019) estimates the probability of cohabitation with a partner (for details, see Table 1). Since the model is specified separately for men and women, the estimated propensities for a relationship might yield an imbalance of men and women available on the partner market. By only considering people in the same district, regional aspects of partner markets and the importance of spatial distance for partnerships are modeled.[8]

In the second step, the selected persons are matched.[9] The probability for cohabitation is estimated with a logit model using German Microcensus data. To account for potential age difference of the partners, spline functions are used in terms of generalized additive models (Wood, 2017). Therefore, also rare but possible partnerships (for example, with large age differences) are generated. The simulated imbalances of regional partner demands models the option of better available choices in asymmetric partner markets (cf. Klein, 2000).

Marriage

Within the marriage module, only single, widowed, or divorced partners are eligible for marriage. The module starts with unmarried couples within existing households. Information on the household and individuals is used to predict the couple's probability to marry. The model is based on data from the German Socio-Economic Panel (for details of the model, cf. Table 1). Given the estimated transition probabilities, marriages are simulated by updating family status for the involved partners.

---

7 The main mechanism is homogamy with respect to age, socio-economic status, and nationality (Klein, 2015).

8 People living in different districts are not matched in the module for two reasons. On the one hand considering all potential partners would result in high computational costs. On the other hand matched partners from different districts would again lead to regional mobility and subsequently distort the marginal distribution resulting from the Regional Mobility module.

9 Due to limited information on same-sex partnerships, the partnership module is restricted to hetero-sexual partnerships only.

Moving into Shared Apartments

Since the number of apartment-sharing communities is likely to be high in large cities and university towns, neglecting this type of housing would lead to an unrealistically high number of single-person households. In the module, only people currently living alone between 18 and 35 years old are eligible to move into a shared apartment.

The probability for moving into shared apartments is based on data from the German Microcensus and is estimated with a logit model including only age and gender as independent variables (cf. Table 1). The estimated probabilities for each district are calibrated via iterative proportional fitting so that the proportion of people living in a shared apartment remains the same after the first simulation period. The proportion of people moving into shared apartments is then left constant, such that a change of the total number of people living in shared apartments is a result of changing population structures.

To form new households, all relocating persons are randomly matched, such that the average household size of three persons is created while distributional size assumptions are approximately met.

School Education

The School Education module consists of two sub-modules, (1) a School Enrollment module and (2) a module for assigning educational qualifications.

The Enrollment module assigns the time of enrollment based on relative frequencies from official data (Statistisches Bundesamt, 2020b). A child is either enrolled early (at age 5), regularly (at age 6) or late (at age 7).[10] The probabilities are available by federal state and sex for different types of primary schools.[11]

The School Education module assigns the duration of school attendance and the resulting certificate. Grade levels of students are promoted yearly. To take repetitions of classes into account, a pragmatic approach was chosen: The duration of primary schooling is estimated for each child. Based on the time of enrollment and the estimated duration, the age at which the students will complete the fourth grade is calculated. At this age, the child will be promoted to the fifth grade. Before this, they are simply in primary school.

10 In Germany, there are different key dates for school enrollment depending on the federal state. In MikroSim we do not take them into account, since we do not model birthdays and we use a yearly time framework to update characteristics.

11 An initial attempt to estimate the probabilities by a model using the National Education Panel Study (NEPS) (Blossfeld, Roßbach, and Maurice, 2011) was discarded due lack of predictive power: The model including education, partner education, age, age of partner, work, marital status, and body length at birth yielded a McFaddens Pseudo-R^2 of 0.03.

Children at the beginning of the simulation must be assigned a grade level. They are assigned a grade level according to their age, not considering any covariates such as repeated classes or late enrollments.

The NEPS (Starting Cohort 2) is used to estimate the probabilities for the duration of primary schooling. The current module promotes people yearly until they leave school. A model predicting re-attendance of a class from fifth grade onward will be implemented in later stages of the project.

Since school qualifications in Germany can be obtained after grade 9, degrees are tentatively assigned at this grade. As soon as a qualification is assigned, the school career is continued depending on this qualification. The probabilities for grades are estimated using a multinomial logit model based on NEPS (Starting Cohort 4) data (see Table 1 for details).

Vocational Training

In the Vocational Training module, people who left school are assigned a job qualification. Possible values are no vocational training, vocational training degree, or university degree.[12] All model estimates are based on the dataset "Growing up in Germany" (AID:A II) produced by the German Youth Institute.

The time until graduation (1–6 years) of people with an assigned Bachelor's degree is taken from official statistics (Statistisches Bundesamt, 2019b). Persons not obtaining a degree receive a vocational training which after 3 years ends with a professional qualification. Future versions of the module will use distributions from official data for the duration of training.

At the start of the simulation, people already attending a university need an estimate of the duration of their attendance, which is estimated using a regression model based on a student survey data (Georg, Ramm, & Bundesministerium für Bildung und Forschung, 2016). Age is used to predict the stage of the academic career of the student.

Employment

The Employment module consists of three models: First, the employment status is assigned to the individuals in the base dataset. Second, full-time and part-time employment are estimated for the working population. Hereby, only individuals between the age of 15 to 66 years are considered. Third, the simulated objects are sent into retirement.

12 Degrees from universities of applied sciences and PhDs are not modeled because the proportions in the population are small and are often coded as university degree in survey data.

The required transition probabilities are estimated by a multinomial logistic model using German Microcensus data. The model uses individual-level variables (such as age, gender, education, and employment status last year), household characteristics (children and type of relationship), and considers the regional variation in labor market status to predict employment status.

For the working subpopulation the probabilities for part-time or full-time employment is estimated using the same predictors (see Table 1).

In the module, currently all people passing the age-threshold of 67 get retired. In a later version we will account for early retirement.

Income

A module of central importance is the Income module. This module provides necessary information for policy analyses, such as tax or family policy reforms. Moreover, income is an important explanatory variable for other models, such as fertility decisions, internal migration, or education opportunities of children.

The Income module is based on the Taxpayer Panel, an administrative data source covering the entire population of taxpayers in Germany from 2001 to 2014. Since the tax data does not contain detailed socio-demographic variables, the dataset will be enhanced by Microcensus data using statistical matching methods (predictive mean matching and nearest-neighbor random hotdeck).

The purpose of the module is the regional prediction of income and its changes over time. Estimates of individual incomes based on a mixed model will be calibrated with published regional data, such as income, poverty, and inequality indicators. To model changes, a two step approach is used. First, the probability of a change in income is predicted, then the differences to the previous year is modeled.

Application Modules

The core simulation model can be extended easily. Currently, two extensions are implemented: (1) labor market outcomes of migrants in Germany and (2) elderly care in the family. Other modules will be added in the future. We describe the current non-core modules briefly.

Labor Market Outcomes for Migrants

The second subject-matter topic is the development of labor market qualifications in the migrant population in Germany. The aim of the module is to supplement the projection of labor market integration with a regional perspective since both the

allocation of migrants and the labor market outcomes of individuals differ regionally.

Membership to different migrant groups (e.g., EU/non-EU) and migrant generations (i.e. born elsewhere/second-generation) are modeled regionally. A specific citizenship model within the module generates naturalization probabilities for all migrants in the simulation. The models specified in Table 1 will be estimated for the migrant population. To estimate the integration of migrants in the labor market, the Employment module and its first sub-module concerning the labor market status of the simulated individuals are of interest.

The relative labor market positioning of migrants in comparison to the majority population is predicted dependent on the assignment of outcomes in the School and Vocational Education modules run previously. The Income module will allow the study of income disparities and their potential change over time for the migrant population. Through the planned Citizenship module and the extensions in the Employment module, we can then estimate different scenarios of regional labor market integration for different ethnic minorities.

Elderly Care

A central subject matter problem in the MikroSim project is the increase of informal care for the elderly depending on demographic changes and new family structures. The current aim of the module is the study of the development of intra-family care. Therefore, household structures have to be updated. For example, children and grandchildren can be potential informal care providers for the elderly. Since the same function can be performed by family members outside the household, a mechanism to add these external families to a household is needed. The details of this mechanism are the subject of ongoing research.

In addition to this modeling of care supply, the demand for care has to be simulated. Therefore, a model for the degree and duration of care required is needed. Since the degrees of care have to be consistent with other variables in the model, additional constraints have to be fulfilled.

Adding this module with complex modeling of both the need for care and the type of care allows then to answer various questions in this area. Besides the possibility to forecast certain parameters such as the number and the proportion of people in need of care under ceteris paribuis conditions, the complex interdependency structure of the MikroSim model also allows to investigate the future effects of various social processes, such as the demographic change or the differentiation of family forms.

An example of this is the discussion between the medicalization or the compression thesis. Using microsimulation methods, for example, allows to analyze whether the progress in curative medicine leads to a higher number of people in

need of care due to a higher amount of years spent in disease (medicalization thesis). In contrast, the compression thesis assumes that the number of years spent in sickness will not increase, but that more of life will be spent in health. Adding this assumption as a complex "what-if" scenario then allows to compare the effects on a variety of target values that are either directly linked to this phenomenon (e.g. number of people in need of care, burden on the health care system), or that result from the interdependence structure of the simulation model (side effects such as a potential change in women's labor force participation due to the need to care for dependents).

Achievements, current work, and further development

After the first phase of work, the datasets required for estimation have been obtained, harmonized, documented and been used to estimate parameters for cross-sectional characteristics and transition probabilities. The base dataset has been updated to new margins and enhanced by survey data estimates.

The basic structure of the simulation model was planned and implemented as separate modules as shown in Figure 1. For each module, the processes required for updating the model were specified and estimated using available data. Resulting estimates were calibrated to known (regional) totals. The program code has been documented and tested. Currently, first test runs of the schooling simulation are in progress.

Within the next period, the refinements described above will be implemented, e.g. rural-urban disparities and improved regional patterns. After that, sensitivity studies and policy scenarios will be run and analyzed. We intend to publish first simulation results 30 months after the project has started. The availability of further data will furnish additional calibration methods. The modules will be fine-tuned further, leading to improved reproduction of known regional patterns and rare subgroups. To support continuous data updates while preserving model reproducibility, a data versioning system will be implemented in the simulation environment.

During the next funding period, additional modules will be implemented to test scenarios for policy studies in housing, health service research and urban travel demand. We intend to open the simulation model for other research groups by building a research data center, operating on similar principles as other data research centers in Germany.

References

Alfons, A., Filzmoser, P., Hulliger, B., Kolb, J.-P., Kraft, S., Münnich, R., & Templ, M. (2011). Synthetic data generation of SILC data (Research Project Report No. 6.2). Advanced Methodology for European Laeken Indicators.

Alfons, A., Kraft, S., Templ, M., & Filzmoser, P. (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, 20 (3), 383–407.

Bækgaard, H. (2002). Micro-macro linkage and the alignment of transition processes: Some issues, techniques and examples. University of Canberra, National Centre for Social and Economic Modelling.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (eds.). (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Sonderheft 14*.

Burgard, J. P., Dieckmann, H., Krause, J., Merkle, H., Münnich, R., Neufang, K. M., Schmaus, S., et al. (2020). A generic business process model for conducting microsimulation studies. *Statistics in Transition New Series*, 21 (4), 191–211.

Burgard, J. P., Krause, J., Merkle, H., Münnich, R., & Schmaus, S. (2019). Conducting a dynamic microsimulation for care research: Data generation, transition probabilities and sensitivity analysis. In A. Steland, E. Rafajłowicz, & O. Okhrin (Eds.), *Workshop on stochastic models, statistics and their application: Dresden, Germany, March 2019* (pp. 269–290). Cham: Springer.

Burgard, J. P., Krause, J., Merkle, H., Münnich, R., & Schmaus, S. (2020). Dynamische Mikrosimulationen zur Analyse und Planung regionaler Versorgungsstrukturen in der Pflege. In M. Hannappel & J. Kopp (Eds.), *Mikrosimulationen* (pp. 283–313). Wiesbaden: Springer VS.

Burgard, J. P., Krause, J., & Schmaus, S. (2020). Estimation of regional transition probabilities for spatial dynamic microsimulations from survey data lacking in regional detail. *Computational Statistics & Data Analysis*, 154.

Destatis. (2020). Bevölkerung: Wanderungen. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Wanderungen/_inhalt.html.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2018a). Mikrozensus 2012, SUF, Version 0. doi:10.21242/12211.2012.00.00.3.1.0.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2018b). Mikrozensus 2013, SUF, Version 0. doi:10.21242/12211.2013.00.00.3.1.0.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder. (2018c). Mikrozensus 2014, SUF, Version 0. doi:10.21242/12211.2014.00.00.3.1.0.

Georg, W., Ramm, M., & Bundesministerium für Bildung und Forschung. (2016). Learning conditions and student orientations 2012/13. doi:10.4232/1.12510.

German Youth Institute. (2014). Growing up in Germany: Everyday life's world (AID:A II). München. Retrieved April 7, 2020, from https://surveys.dji.de/index.php?m=msw,0&sID=107.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German socio-economic panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, 239 (2), 345–360. doi:10.1515/jbnst-2018-0022.

Hannappel, M., & Kopp, J. (Eds.). (2020). *Mikrosimulationen: Methodische Grundlagen und ausgewählte Anwendungsfelder*. Wiesbaden: Springer VS.

Huang, Z., & Williamson, P. (2001). A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata (Working Paper No. 2001/2). Department of Geography, University of Liverpool. Liverpool.

Klein, T. (2000). Partnerwahl zwischen sozialstrukturellen Vorgaben und individueller Entscheidungsautonomie. *Zeitschrift für Soziologie der Erziehung und Sozialisation*, 20 (3), 229–243.

Klein, T. (2015). Partnerwahl. In P. B. Hill & J. Kopp (Eds.), *Handbuch Familiensoziologie* (pp. 321–343). Wiesbaden: Springer.

Klevmarken, A. (2008). Dynamic microsimulation for policy analysis: Problems and solutions. In A. Klevmarken & B. Lindgren (Eds.), *Simulating an ageing population: A microsimulation approach applied to Sweden*. Bingeley: Emerald Group Publishing Limited.

Kolb, J.-P. (2013). Methoden zur Erzeugung synthetischer Simulationsgesamtheiten (Doctoral dissertation, University of Trier, Trier). Retrieved from https://ubt.opus.hbz-nrw.de/files/590/Diss_Kolb_JP.pdf.

Laarhoven, P. J. M., & Aarts, E. H. L. (1987). *Simulated annealing: Theory and applications*. Dordrecht: Springer.

Li, J., & O'Donoghue, C. (2013). A survey of dynamic microsimulation models: Uses, model structure and methodology. *International Journal of Microsimulation*, 6 (2), 3–55. doi:10.34196/ijm.00082

Li, J., & O'Donoghue, C. (2014). Evaluating binary alignment methods in microsimulation models. *Journal of Artificial Societies and Social Simulation*, 17 (1), 15. doi:10. 18564/jasss.2334

Merz, J. (1991). Microsimulation—a survey of principles, developments and applications. *International Journal of Forecasting*, 7 (1), 77–104.

Münnich, R., Burgard, P., J., & Vogt, M. (2013). Small Area-Statistik: Methoden und Anwendungen. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 6 (3/4), 149–191. doi:10.1007/s11943-013-0126-1

Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P., & Kolb, J.-P. (2012). Stichprobenoptimierung und Schätzung im Zensus 2011. *Statistik und Wissenschaft*. Wiesbaden: Statistisches Bundesamt.

Münnich, R., Schnell, R., Kopp, J., Stein, P., Zwick, M., Dräger, S., Merkle, H., Obersneider, M., Richter, N., Schmaus, S. (2020). Zur Entwicklung eines kleinräumigen und sektorenübergreifenden Mikrosimulationsmodells für Deutschland. In M. Hannappel & J. Kopp (Eds.), *Mikrosimulationen* (pp. 109–140). Wiesbaden: Springer VS.

Münnich, R., & Schürle, J. (2003). On the simulation of complex universes in the case of applying the German Microcensus (DACSEIS Research Paper Series No. 4). Eberhard Karls University of Tübingen.

O'Donoghue, C., & Dekkers, G. (2018). Increasing the impact of dynamic microsimulation modelling. *The International Journal of Microsimulation*, 11 (1), 61–96.

Orcutt, G. H. (1957). A new type of socio-economic system. *The Review of Economics and Statistics*, 39 (2), 116–123.

Perese, K. (2002). Mate matching for microsimulation models (Technical Report No. 3). Congressional Budget Office. Washington. Retrieved from https://www.cbo.gov/publication/14211.

Rahman, A., & Harding, A. (2016). *Small area estimation and microsimulation modeling*. Boca Raton: CRC Press.

Rao, J., & Molina, I. (2015). *Small area estimation, 2nd edition*. Wiley Series in Survey Methodology. Hoboken: John Wiley and Sons, Ltd.

Schneider, N. F. (2015). Familie in Westeuropa: Von der Institution zur Lebensform. In P. B. Hill & J. Kopp (Eds.), *Handbuch Familiensoziologie* (pp. 21–54). Wiesbaden: Springer.

Schnell, R., & Handke, T. (2020). Neuere bevölkerungsbezogene Mikrosimulationen in Großbritannien und Deutschland. In M. Hannappel & J. Kopp (Eds.), *Mikrosimulationen* (pp. 35–56). doi:10.1007/978-3-658-23702-8_3

Statistisches Bundesamt. (2019a). Bevölkerung: Eheschließungen, Geborene und Gestorbene 2018 nach Kreisen. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Eheschliessungen-Ehescheidungen-Lebenspartnerschaften/Publikationen/Downloads-Eheschliessungen/eheschliessungen-geborene-gestorbene-5126001187004.html.

Statistisches Bundesamt. (2019b). Bildung und Kultur: Prüfungen an Hochschulen. Retrieved January 21, 2020, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bildung-Forschung-Kultur/Hochschulen/Publikationen/Downloads-Hochschulen/pruefungen-hochschulen-2110420187004.pdf.

Statistisches Bundesamt. (2020a). Bevölkerung: Sterbefälle und Lebenserwartung. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Sterbefaelle-Lebenserwartung/_inhalt.html.

Statistisches Bundesamt. (2020b). Schulanfänger: Bundesländer, Schuljahr, Geschlecht, Einschulungsart, Schulart. Retrieved April 7, 2020, from https://www-genesis.destatis.de/genesis/online?operation=table&code=21111-0010.

Stein, P., & Bekalarczyk, D. (2016). Zur Prognose beruflicher Positionierung von Migranten der dritten Generation. In R. Bachleitner, M. Weichbold, & M. Pausch (Eds.), *Empirische Prognoseverfahren in den Sozialwissenschaften* (pp. 223–257). Wiesbaden: Springer.

Stephensen, P. (2016). Logit scaling: A general method for alignment in microsimulation models. *International Journal of Microsimulation*, 9 (3), 86–102.

Sutherland, H. (2018). Quality assessment of microsimulation models: The case of Euromod. *International Journal of Microsimulation*, 11 (1), 198–223.

Tanton, R. (2014). A review of spatial microsimulation methods. *International Journal of Microsimulation*, 7 (1), 4–25.

van Imhoff, E., & Post, W. (1998). Microsimulation methods for population projection. *Population: An English Selection*, 10 (1), 97–138.

Williamson, P. (2012). An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation. In R. Tanton & K. L. Edwards (Eds.), *Spatial microsimulation: A reference guide for users* (pp. 19–47). Dordrecht: Springer.

Wood, S. N. (2017). *Generalized additive models: An introduction with R (2nd ed.)*. Boca Raton: CRC Press.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In 88th annual meeting of the transportation research board.

Zinn, S. (2012). A mate-matching algorithm for continuous-time microsimulation models. *International Journal of Microsimulation*, 5 (1), 31–51.

Zwick, M., & Emmenegger, J. (2020). Mikrosimulation und Gesellschaftspolitik – ein kurzer historischer Abriss. In M. Hannappel & J. Kopp (Eds.), *Mikrosimulationen* (pp. 17–34). doi:10.1007/978-3-658-23702-8_2

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)